An Analysis of Structured Optimal Policies for Hypertension Treatment Planning: The Tradeoff Between Optimality and Interpretability

Gian-Gabriel P. Garcia[1], Lauren N. Steimle[1], Wesley J. Marrero[2], and Jeremy B. Sussman[3]
[1]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA
[2]Institute for Technology Assessment, Massachusetts General Hospital, Harvard Medical School, Boston, MA
[3]Department of Internal Medicine, Michigan Medicine, University of Michigan, Ann Arbor, MI

Problem definition: In medical decision-making, Markov decision processes (MDPs) are useful for deriving optimal treatment policies when a patient's health evolves stochastically over time. Yet, optimal policies may lack structure that is interpretable to human decision-makers. When interpretability is valued by practitioners, suboptimal yet interpretable policies may be preferred over uninterpretable optimal policies. Interpretability is especially critical in hypertension treatment, where complicated clinical guidelines have drawn substantial controversy from practitioners. In this research, we design and analyze a class of interpretable policies for MDPs which leverage the natural interpretability of monotonicity, i.e., the intensity of the prescribed action increases with the severity of the state.

Methodology/results: We present mixed integer programming formulations to obtain the optimal monotone policy and class-ordered monotone policy (CMP). The novel CMP generalizes monotone policies by imposing monotonicity over classes of states and actions rather than states and actions themselves. We show that the optimal monotone policy can be found in polynomial time, but the degree is non-trivial. We then analyze the performance gap of monotone policies and CMPs relative to the optimal policy using the price of interpretability (PI). Under mild conditions, we prove that the CMP achieves a PI no greater than the optimal monotone policy. Finally, we demonstrate the practicality of these methods for hypertension treatment and derive patient-level and population-level insights. Overall, the CMP's flexibility allows it to outperform the monotone policy, achieving the greatest benefit for patients with stage 2 hypertension. Nevertheless, both interpretable policies retain low PIs and save over 3,200 quality-adjusted life years and prevent nearly 300 cardiovascular events over clinical guidelines while retaining a clinically intuitive structure.

Managerial implications: CMPs retain the interpretability of monotone policies while achieving superior performance. In practical applications, CMPs can provide more instinctive strategies than the optimal policy with minimal losses in performance.

Key words: Markov decision processes, healthcare applications, medical decision-making, interpretability, cardiovascular disease, personalized treatment planning

1. Introduction

Markov decision processes (MDPs) are a commonly used mathematical framework for analyzing sequential decision-making processes in which a decision maker (DM) aims to control a system that evolves stochastically over time. MDPs have been applied to many domains including healthcare, energy, and finance (Puterman 2014). In some cases, the optimal decision rules generated by MDPs are naturally interpretable. Here, *interpretable policies* are decision-making policies for which the mapping between the state of the system and its corresponding decision is amenable to human intuition, cognition, and pattern recognition. Naturally, this definition implies that the class of interpretable policies may change from context to context.

Monotone policies are often considered to be interpretable policies because they implement actions that are non-decreasing (or non-increasing) with respect to the system state. For instance, in medical decision-making, an interpretable policy might recommend more severe treatment options for patients who are sicker. Because such interpretable policies may facilitate implementation in practice, it is often desirable to show that an optimal policy also happens to be monotone.

Unfortunately, the sufficient conditions that guarantee the existence of an optimal policy that is monotone are often violated to some degree. In these situations, a DM who values interpretability may be faced with the following questions: Should I solve the MDP and then adjust the policy to be interpretable? Should I solve a slightly perturbed version of the MDP which satisfies the sufficient conditions to guarantee an interpretable policy? What is the cost of using the best interpretable policy instead of the actual optimal policy? These questions demonstrate that, although MDPs can be used to optimize sequential decision-making in theory, there still exist barriers to implementing these optimal policies in practice. In this research, our goal is to design interpretable policies for MDPs by leveraging the natural interpretability of monotone policies and assess the "price" of using this type of policy instead of the optimal policy.

1.1. MDPs in Medical Decision-Making

This research is motivated by medical decision-making, a setting in which MDPs are often applied. Surveys by Denton et al. (2011), Capan et al. (2017), Saville et al. (2018), and Chanchaichujit et al.

(2019) comprehensively review many MDPs in the extant literature. More recent examples include papers by Chen et al. (2018), Hicklin et al. (2018), Suen et al. (2018), Agnihothri et al. (2018), Lee et al. (2018), Ayer et al. (2019), Boloori et al. (2020), Skandari and Shechter (2021) and Marrero et al. (2021). While MDPs are useful for modeling complex decision-making in medical settings, their policies may not be implemented in practice due to limited interpretability resulting from the complexity of the policy or a disconnect between the policy and clinician's intuition.

An important clinical setting where DMs may benefit from interpretable medical decision-making policies is the management of atherosclerotic cardiovascular disease (ASCVD). ASCVD, which constitutes coronary heart disease (CHD) and stroke, is among the leading causes of death in the US (Kochanek et al. 2019). Recent reports show CHD and stroke account for 42.6% and 17.0% of deaths due to cardiovascular diseases in the US, respectively (Virani et al. 2020). A major risk factor for ASCVD is patients' blood pressure (BP). Several clinical practice guidelines (Chobanian et al. 2003, James et al. 2014, Williams et al. 2018, Whelton et al. 2018) as well as optimal decision models (Denton et al. 2009, Kurt et al. 2011, Mason et al. 2014, Schell et al. 2016, Steimle et al. 2021) have been developed to manage patients with high BP or hypertension. However, the clinical guidelines may be deemed as subjective (Cohen and Townsend 2018, Solberg and Miller 2018) and the policies obtained with MDPs may be complex and not easily interpretable (Lakkaraju and Rudin 2017). With the goal of increasing the acceptability of MDPs in medical practice, this research extends these prior works by developing interpretable treatment strategies for the personalized management of hypertension.

1.2. Interpretable Policies for MDPs

Early work on interpretability for MDPs has focused on structured policies that are optimal for specific applications of MDPs (e.g., inventory control (Bellman et al. 1955, Schäl 1976)) and on monotone policies (Serfozo 1976), where optimal policies prescribe actions which are monotone in the system state. Monotone policies are considered interpretable in settings where the states and actions are both ordered. Previous research has established sufficient conditions on the MDP

parameters which guarantee the existence of an optimal policy that is monotone (Smith and McCardle 2002, Puterman 2014, §6.11). Our research builds on this past work by developing methods to derive an optimal monotone policy even when these sufficient conditions are not satisfied. Moreover, we consider a generalization of the monotone policy, i.e., the *class-ordered monotone policy*, which can be interpretable when the states and/or actions do not follow a strict ordering.

Interpretability has also been of interest in partially-observable MDPs (POMDPs). Monotone policies for POMDPs are described in Lovejoy (1987). There has been some investigation of interpretable and implementable, yet potentially suboptimal, policies for POMDPs. Early work in this area dates back to Littman (1994) and Vlassis et al. (2012) who studied the class of *memoryless policies* for POMDPs. Although memoryless policies are not guaranteed to be optimal for POMDPs, they are interpretable in the sense that the action taken by the DMs depends only on the most recent observation, rather than the entire history of observations and actions. More recently, Chen et al. (2018) and Cevik et al. (2018) describe interpretable policies for POMDPs in the context of cancer screening. Chen et al. (2018) consider a special form of interpretable policies, called "M-switch," which enforces that screening schedules must be at regular intervals and the length of these intervals can only switch M times. The "M-switch" policies are interpretable relative to traditional recommendations from POMDPs for cancer screening in which the optimal policy is not guaranteed to be of a regular frequency. Likewise, Cevik et al. (2018) consider the problem of breast cancer screening under resource constraints and design policies with the property that if it is optimal to screen a patient with a certain risk for developing breast cancer, then it should also be optimal to screen any patient with greater risk. While our research focuses on interpretable policies for completely observable MDPs, we demonstrate that the policy imposed by Cevik et al. (2018) is a special case of the class-ordered monotone policy (see §2.3).

Perhaps the most closely related works to ours is that of Petrik and Luss (2016) and Serin and Kulkarni (1995) who consider interpretable policies in fully observed MDPs. Petrik and Luss (2016) and Serin and Kulkarni (1995) consider interpretable policies for MDPs by first partitioning

the states space of an MDP into K sets. A policy is considered interpretable if the probability of taking an action is the same for all states in the same set. The DM's goal in this setting is to find the best policy among this type of interpretable policy. Serin and Kulkarni (1995) show that this problem is a special case of finding the best memoryless policy in a POMDP. They also prove that in general, there is no guarantee that a deterministic interpretable policy will be optimal. Further, they showed that the optimal policy depends on the initial distribution over the states and proposed an iterative method for finding local optimal solutions. Petrik and Luss (2016) later showed that solving for randomized or deterministic interpretable policies is NP-hard and proposed a mixed-integer program (MIP) to solve this problem. In this article, we propose a more general form of interpretable policy wherein both the state space and action space are partitioned into classes. We also consider monotonicity requirements as defined on the space of state classes and action classes adding to the interpretability of our approach. We argue that this type of policy is interpretable while also achieving better performance than monotone policies.

1.3. Contributions

In this article, we develop and analyze new methods for designing interpretable policies for MDPs. This research makes the following contributions:

1. We introduce a new type of interpretable policy for MDPs, which we call a *class-ordered monotone policy (CMP)*. The CMP generalizes many types of interpretable policies, including monotone policies, and we show that under mild conditions on the state and action classes, the optimal CMP is guaranteed to perform no worse than the optimal monotone policy.
2. We introduce the *price of interpretability (PI)* in MDPs, which measures the difference between the optimal value of the MDP and the value corresponding to the best interpretable policy. This metric can guide DMs who value interpretability and wish to know the cost of using the best interpretable policy instead of the optimal policy.
3. We provide exact solution methods for finding optimal monotone policies and CMPs. Our exact solution methods are MIP formulations which use logic-based constraints to enforce

(class-ordered) monotonicity. We also provide heuristic methods to derive initial feasible solutions to these MIPs, which can decrease the solution time via warm starting.

4. We demonstrate the practicality of our approach in a case study where an MDP is used for personalized hypertension treatment. Our study demonstrates that while an optimal policy may not be monotone and that the best monotone policy may not be sufficiently close to optimal, the optimal CMPs can achieve both interpretability and high performance, i.e., it pays a low PI. Moreover, the CMP's performance is robust to different state and action class definitions — implying that the CMP is amenable to context-specific interpretability.

The remainder of this article is organized as follows. In §2, we introduce and analyze monotone policies and CMPs, as well as the PI. In §3, we use an MDP to derive hypertension treatment plans for the prevention of ASCVD and analyze the PI for monotone policies and CMPs. Finally, in §4, we conclude with a discussion of our findings and directions for future research.

2. Methodology

In this section, we first present our basic problem setting of infinite horizon MDPs. The infinite horizon setting allows us to convey all of the key ideas in our methods and analysis. Next, we formally define a monotone policy for this class of MDPs and provide an MIP formulation to determine the optimal monotone policy. Then, we define the notion of a CMP and show how to extend our MIP formulation to determine the optimal CMPs. Finally, we define the PI and analyze the CMP and optimal monotone policy with respect to the PI. Proofs for all technical results are provided in §EC.1 of the e-companion. While we develop our methods in the infinite horizon context, these methods be flexibly adapted to finite horizon MDPs by adding a temporal component. Doing so also allows for policies that impose monotonicity on the time in addition to the state. We take this approach in our case study (see §3) and detail these modifications in the e-companion (see §EC.4.1).

2.1. Problem Setting

We consider infinite horizon MDPs which are used to model sequential decision-making in uncertain environments. The underlying Markov chain is defined by a set of states, $\mathcal{S} = \{1, \ldots, S\}$. At each

decision epoch $t \in \mathcal{T} = \{0, 1, \ldots\}$, the DM observes the state $s \in \mathcal{S}$ and then performs an action $a \in \mathcal{A} = \{1, \ldots, A\}$. When action a is performed in state s, the DM receives a finite reward $r(s,a)$ and the Markov chain transitions to a new state s' according to a transition probability matrix P with entries $P(s'|s,a) = \mathbb{P}(s_{t+1} = s'|s_t = s, a_t = a)$. Rewards are discounted at a rate $\gamma \in (0,1)$. At the first decision epoch $t = 0$, the state of the system is probabilistically generated by an initial state distribution $\alpha \in \{\mathbb{R}_+^S : \sum_{s' \in \mathcal{S}} \alpha(s) = 1\}$. This MDP is summarized by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, P, r, \alpha)$.

The DM aims to select actions in order to maximize the expected total discounted rewards over the planning horizon. A deterministic stationary policy $\pi : \mathcal{S} \to \mathcal{A}$ is a decision rule which maps each state to an action. We denote by Π the set of all admissible policies. The total discounted reward accrued by an MDP given a policy $\pi \in \Pi$ and initial state distribution α is given by

$$J^\pi(\alpha) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(s_t)) \middle| \mathbb{P}(s_0 = s) = \alpha(s) \right].$$

The optimal policy is given by $\pi^* = \arg\max_{\pi \in \Pi} J^\pi(\alpha)$. It is well-known (Puterman 2014, Ch. 6.9) that there exists a deterministic and stationary policy that is optimal and that the value of $J^{\pi^*}(\alpha)$ can be obtained by solving a linear program (LP):

$$\text{(MDP-LP)} \quad J^{\pi^*}(\alpha) = \min_{\mathbf{v}} \quad \sum_{s \in \mathcal{S}} \alpha(s) v(s) \tag{1a}$$

$$\text{subject to:} \quad v(s) \geq r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) v(s') \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}, \tag{1b}$$

where $\mathbf{v} \coloneqq [v(s)]_{s \in \mathcal{S}}$ represents the vector of state value functions. In MDP-LP, the parameter $\alpha(s) > 0$ for all $s \in \mathcal{S}$. Moreover, for each $s \in \mathcal{S}$, $\pi^*(s)$ is equal to the action where (1b) is tight. Note that π^* is independent of α due to the principle of optimality.

2.2. Monotone Policies

In this section, we assume that the sets $\mathcal{S}$ and $\mathcal{A}$ are ordered. We now restrict our attention to the set of monotone policies, denoted by Π^M, and defined as follows.

Definition 1 (Monotone Policy). Under ordered states and actions, a monotone policy is a policy $\pi : \mathcal{S} \to \mathcal{A}$ such that $\pi(s) \leq \pi(s')$ for all $s, s' \in \mathcal{S}$ such that $s \leq s'$.

Monotone policies are appealing to practitioners because they can be more easily interpreted and implemented compared to optimal policies may lack structure. For instance, physicians may find treatment strategies more interpretable if the policies follow a natural order, such as increasing treatment intensity on the severity of a patient's health condition. In the current literature, researchers typically identify sufficient conditions on the MDP data (i.e., P and r) which guarantee that there exists an optimal policy which is monotone, i.e., $\pi^* \in \Pi^M$. In contrast, our aim is to determine the policy $\pi^M \in \Pi^M$ which achieves the greatest total discounted reward without requiring any conditions on the MDP data. Formally, our optimization problem is given by:

$$\pi^M = \arg\max_{\pi \in \Pi^M} J^\pi(\alpha). \tag{2}$$

To determine π^M, we modify MDP-LP by adding binary decision variables to impose the desired monotone structure within the policy. We formulate this Monotone MDP MIP (M-MIP) as:

$$\begin{aligned}
\text{(M-MIP)} \quad J^{\pi^M}(\alpha) = \max_{\mathbf{v}} \quad & \sum_{s \in \mathcal{S}} \alpha(s) v(s) && \text{(3a)}\\
\text{subject to:} \quad & v(s) \le r(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) v(s') + M_{s,a}(1 - x(s,a)) \\
& \qquad \text{for all } s \in \mathcal{S},\ a \in \mathcal{A} && \text{(3b)}\\
& \sum_{a \in \mathcal{A}} x(s,a) = 1 \text{ for all } s \in \mathcal{S} && \text{(3c)}\\
& x(s,a) \le \sum_{a' \ge a} x(s+1, a') \text{ for all } s \in \mathcal{S} \setminus S, a \in \mathcal{A} && \text{(3d)}\\
& x(s,a) \in \{0,1\} \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}. && \text{(3e)}
\end{aligned}$$

In M-MIP, we highlight that (1a) and (3a) differ because we use a slightly different convention for the epigraph constraints. That is, we modify (1b) with binary variables and the big-M method to obtain (3b). Generally, big-M constraints are known to weaken MIP formulations. Letting $v^*(s)$ represent the optimal value function for state s (which can be obtained by solving (1)), we strengthen our formulation by setting $M_{s,a} = v^*(s)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$ since the left-hand side of (3b) will never exceed $v^*(s)$. The remaining constraints (3c) and (3d) ensure that the resulting policy is

deterministic and monotone in s, respectively. Taking $\pi^M(s) = \arg\max_{a \in \mathcal{A}} x(s,a)$ gives the optimal monotone action for each state. In comparison to the optimal policy π^*, the optimal monotone policy π^M depends on the initial state distribution α.

PROPOSITION 1. *The optimal monotone policy π^M depends on the initial distribution α.*

Proposition 1 implies that optimal monotone policies, in general, are history-dependent. Practically speaking, this result implies that the initial state distribution must be known to create an optimal monotone policy. From a technical perspective, this result implies that traditional "off the shelf" algorithms for solving MDPs which do not account for the initial distribution (e.g., value iteration and policy iteration) may not work for determining optimal monotone policies. Hence, these algorithms may need to be modified for this class of policies. Despite the need for these modifications, Theorem 1 shows that we can still find π^M in polynomial time.

THEOREM 1. *The optimization problem in* (2) *is solvable in polynomial time.*

We prove this result by first showing that the number of monotone policies grows as a polynomial in the number of states and the number of actions in the following Lemma and Corollary.

LEMMA 1. *The number of monotone policies is given by* $\binom{S+A-1}{A-1}$.

As an immediate consequence of Lemma 1, we have the following Corollary:

COROLLARY 1. *The number of monotone policies grows as a polynomial in S and A.*

Theorem 1 follows from Corollary 1 because there are a polynomial number of policies to evaluate and the fact that each policy can be evaluated in polynomial time by solving an LP. Despite this result, the degree of this polynomial is nontrivial; approximation algorithms may still be needed for MDPs with large state and action spaces.

Although monotone policies are highly desirable due to their interpretability, they require the states and actions to follow a strict ordering. However, in practice, a strict ordering across states and actions may be difficult to define, especially for multi-dimensional state and actions spaces. Thus, the DM may feel more comfortable by defining ordering on groups of states and groups of actions. We address this limitation in §2.3 by introducing a new form of interpretable policy.

	Action-class Mapping		
State-class Mapping	$\Psi(a) = a, \forall a \in \mathcal{A}$	Ψ general	$\Psi(a) = \Psi(a'), \forall a, a' \in \mathcal{A}$
$\Theta(s) = s, \forall s \in \mathcal{S}$	Monotone policies	CMP	CMP
Θ general	CMP	CMP	CMP
$\Theta(s) = \Theta(s'), \forall s, s' \in \mathcal{S}$	CMP	CMP	General policies

Table 1 A characterization of policies by type of state-class mapping Θ and action-class mapping Ψ. CMP = class-ordered monotone policy

2.3. Class-ordered Monotone Policies

The *class-ordered monotone policy (CMP)* generalizes monotone policies where monotonicity constraints hold on ordered classes of states and actions rather than the states and actions themselves. Specifically, suppose that $\mathcal{S}$ is partitioned into ordered state-classes $\mathcal{S}_1, ..., \mathcal{S}_K$ indexed by the set $\mathcal{K} = \{1, \ldots, K\}$ and $\mathcal{A}$ is partitioned into ordered action-classes $\mathcal{A}_1, ..., \mathcal{A}_G$ indexed by the set $\mathcal{G} = \{1, \ldots, G\}$. Each state is mapped to exactly one state class through the function $\Theta : \mathcal{S} \to \mathcal{K}$ and each action is mapped to exactly one action class through the function $\Psi : \mathcal{A} \to \mathcal{G}$. The state classes can be interpreted such that for any $k' > k$, any state in class $S_{k'}$ is "more severe" than any state in class S_k (and a similar interpretation can be used for action classes). States and actions within a class are not required to be ordered. Using this construction, we now define CMPs.

Definition 2 (Class-ordered monotone policy). A policy π is a *CMP* if $\Theta(s) \geq \Theta(s')$ implies $\Psi(\pi(s)) \geq \Psi(\pi(s'))$.

While CMPs do not enforce strict monotonicity across states and actions, they retain the natural interpretability inherent in monotone policies. In fact, CMPs are a generalization of monotone policies. As we will show later in this section, CMPs also have desirable properties in terms of their performance. Table 1 characterizes the relationship between monotone policies, CMPs, and generalized policies in terms of the functions Θ and Ψ.

Let $\Pi^{CM}_{\Theta,\Psi}$ be the set of CMPs with respect Θ and Ψ. The optimal CMP is the solution to:

$$\pi^{CM} = \arg\max_{\pi \in \Pi^{CM}_{\Theta,\Psi}} J^{\pi}(\alpha). \tag{4}$$

Note that, in general, π^{CM} is dependent on the initial state distribution α (see Remark EC.1 in §EC.1 of the e-companion). To solve (4), we modify M-MIP by replacing constraint (3d) with the following set of constraints:

$$\sum_{a\in\mathcal{A}_g} x(s,a) \leq \sum_{a'\geq \min \mathcal{A}_g} x(s',a') \text{ for all } s\in\mathcal{S}_k, \quad s'\in\mathcal{S}_{k+1}, \quad k=1,...,K-1, \quad g=1,...,G. \tag{5}$$

With this modification, we define the Class-ordered Monotone MDP MIP (CM-MIP) as

$$\text{(CM-MIP)} \quad J^{\pi^{CM}}(\alpha) = \max_{\mathbf{v}} \quad \sum_{s\in\mathcal{S}} \alpha(s)v(s) \quad \text{subject to:} \quad \text{(3b)-(3c), (3e), (5).} \tag{6}$$

In auxiliary numerical experiments, we found that the dual formulations of M-MIP and CM-MIP can often be solved more quickly than their primal formulations, especially when a warm start solution is provided. We provide details of these dual formulations and the procedure for generating an initial feasible solution in §EC.2 and §EC.3 of the e-companion, respectively.

2.3.1. Examples of CMPs We briefly highlight examples of interpretable policies in previous research which can be classified as a special case of CMPs. Petrik and Luss (2016) and Serin and Kulkarni (1995) (see §1.2) consider classes of interpretable policies wherein all states in a particular class must follow the same action. These policies are an example of CMPs with general Θ and $\Psi(a)=a$ for all $a\in\mathcal{A}$ (see Table 1), although they do not require monotonicity in states. Additionally, Cevik et al. (2018) consider the problem of optimal breast cancer screening under resource constraints. They model the problem as a POMDP and consider a class of policies such that if it is optimal to screen for a patient with a given risk for breast cancer, then it is optimal to screen for any patient whose risk is greater. This policy is an example of the case with general Θ and $\Psi(a)=a$ for all $a\in\mathcal{A}$ (see Table 1, with monotonicity enforced in the states).

2.4. The Price of Interpretability

For both monotone policies and CMPs, the DM may be willing to sacrifice some performance in order to implement an interpretable policy. In these situations, the DM may ask, "what is the cost of implementing the best interpretable policy instead of the best overall policy"? To address this question, we introduce the *price of interpretability (PI)*:

DEFINITION 3 (PRICE OF INTERPRETABILITY). Let $\Pi^I \subset \Pi$ be a specific class of *interpretable* policies. The PI for the policy class Π^I is defined as $\text{PI}(\Pi^I) := \max_{\pi\in\Pi} J^\pi(\alpha) - \max_{\pi\in\Pi^I} J^\pi(\alpha)$. The PI informs the DM about the cost of implementing a policy from the interpretable policy class Π^I rather than implementing the actual optimal policy. To facilitate our analysis of the PI for optimal CMPs relative to optimal monotone policies, we make the following assumption:

ASSUMPTION 1. *The class functions Θ and Ψ are non-decreasing.*

Assumption 1 restricts our attention to the class of CMPs with respect to the strict ordering imposed in standard monotone policies. These conditions are natural in cases where a strict ordering on the state and action spaces can be constructed, but some ordering relations may be questionable and thus, a partial ordering may be more appropriate. For example, research suggests that the benefit of antihypertensive treatment is mainly determined by their BP reduction with little effect attributable to drug-specific factors (Law et al. 2009). While it may be reasonable to order treatment choices in terms of number of medications, it is less clear how antihypertensive drug types should be ordered given the same number of medications. Our analysis begins with the following preliminary result, which establishes useful properties for the class functions Θ and Ψ.

LEMMA 2. *Consider any class functions Θ, Ψ which satisfy Assumption 1.*

1. *Suppose that Θ admits at least two state classes. Let Θ' be a new state class function that merges two classes admitted by Θ, i.e., $\Theta'(s) = \Theta'(s') = k$ for all s, s' where $\Theta(s) = k$ and $\Theta(s') = k+1$ for an arbitrary $k \in \mathcal{K} \setminus \{K\}$. Then $\pi \in \Pi^{CM}_{\Theta,\Psi}$ implies $\pi \in \Pi^{CM}_{\Theta',\Psi}$.*
2. *Suppose that Ψ admits at least two action classes. Let Ψ' be a new action class function that merges two classes admitted by Ψ. Then, $\pi \in \Pi^{CM}_{\Theta,\Psi}$ implies $\pi \in \Pi^{CM}_{\Theta,\Psi'}$*

Lemma 2 shows that any CMP defined by Θ and Ψ remains a CMP under Θ' and Ψ' if Θ' and Ψ' are formed by merging adjacent classes. In Theorem 2, we use these properties to relate the performance of optimal monotone policies with optimal CMPs.

THEOREM 2. *If Θ and Ψ satisfy Assumption 1, we have $\max_{\pi\in\Pi^M} J^\pi(\alpha) \leq \max_{\pi\in\Pi^{CM}_{\Theta,\Psi}} J^\pi(\alpha)$.*

Theorem 2 implies that any monotone relaxation of the state or action ordering through the use of classes will not reduce the value function. Hence, practitioners would be well-served by any reasonable partial ordering (based on the problem context) across states and actions. Furthermore, we can readily state the relationship between monotone policies and CMPs with regard to the PI.

COROLLARY 2. *If* Θ *and* Ψ *satisfy Assumption 1, we have* $PI(\Pi^M) \geq PI(\Pi^{CM}_{\Theta,\Psi})$.

From Theorem 2, it directly follows that the PI will be higher for optimal monotone policies relative to optimal CMPs under the conditions stated in Assumption 1.

REMARK 1. Consider the interpretable policies proposed in Petrik, denoted Π^P with state-classes described by Θ. For those interpretable policies in which Θ satisfies Assumption 1, we have that any arbitrary action-class mapping Ψ that satisfies Assumption 1 is guaranteed to perform at least as well as those in Π^P. That is, $\text{PI}(\Pi^P) \geq \text{PI}(\Pi^{CM}_{\Theta,\Psi})$.

3. Case Study: Personalized Hypertension Treatment

We now apply the optimal monotone policy (π^M) and the optimal CMP (π^{CM}) to the management of ASCVD (i.e., atherosclerotic cardiovascular disease). We begin by providing background on the treatment of hypertension (i.e., high BP), a major risk factor for ASCVD. Then, we describe our MDP, model parameters, and data sources. We present the treatment plans and health outcomes of patients following the optimal policy (π^*), π^{CM}, and π^M, as well as the current clinical guidelines (π^G). Finally, we discuss the implications of the interpretable hypertension treatment plans at a patient and a population level.

3.1. Background on Hypertension Treatment

Following the definition from the 2017 Hypertension Clinical Practice Guidelines, 45.6% of adults in the US have high BP (Whelton et al. 2018). These guidelines define stage 1 hypertension as systolic blood pressure (SBP) of 130-139 mm Hg or diastolic blood pressure (DBP) of 80-89 mm Hg, and stage 2 hypertension as an SBP of at least 140 mm Hg or a DBP of at least 90 mm Hg. The guidelines provide non-pharmacological and pharmacological suggestions for patients with

hypertension and elevated BP, defined as an SBP of 120-129 mm Hg and a DBP smaller than 80 mm Hg. During this case study, we focus on the pharmacological recommendations.

While deriving treatment strategies according to clinical guidelines or based on MDPs can be beneficial, there may be drawbacks associated with each method. For example, the 2017 Hypertension Clinical Practice Guidelines have generated substantial controversy among practitioners (Ioannidis 2018). Medical experts have categorized these guidelines as convoluted and complicated (Cohen and Townsend 2018). On the other hand, the results obtained with MDPs may be hard to implement in practice and to communicate with patients. In this case study, we aim to improve the usability and acceptance of MDPs in clinical settings by providing easily interpretable policies.

3.2. Markov Decision Process Formulation

Since the risk for ASCVD events is nonstationary with respect to patients' age, we model the process of sequentially determining antihypertensive medications as a finite-horizon MDP. We build upon the MDP formulation in Schell et al. (2016) to derive our treatment strategies (i.e., π^*, π^{CM}, and π^M). Our objective is to determine the policies that maximize the expected discounted quality-adjusted life years (QALYs).

The adaptation of our formulations to finite-horizon MDPs for the management of hypertension is described in §EC.4.1 of the e-companion. We add an index t to states, actions, transition probabilities, and rewards to highlight their dependence on the decision epoch, which represents the effect of patients' age on the MDP parameters. As clinicians rarely decrease or discontinue the use of antihypertensive medications to control their patients' blood pressure (Van Der Wardt et al. 2017), we guarantee nondecreasing actions over time by incorporating a temporal component in our state definitions (see §EC.4.2 of the e-companion). The elements of our MDP are as follows:

- $\mathcal{T}'$: planning horizon of 10 years; $\mathcal{T}' \coloneqq \{0, 1, \ldots, T\}$. Decision epoch $t \in \mathcal{T}'$ represents the year $[t, t+1)$ and $T-1$ is the year at which physicians select the last action. We use $T = 10$ to represent the effects of treatment on patients' lifetime. This planning horizon is selected based on conversations with our clinical collaborators and following the major clinical guidelines for the management of cardiovascular diseases (Whelton et al. 2018).

- $\mathcal{S}$: state space comprising patients' demographic information d_t, clinical observations c_t, and health condition h_t. Patients' health condition $h_t \in \{1, \ldots, 10\}$ accounts for their experience with adverse events. The health condition of each patient is one of the following mutually-exclusive categories: healthy ($h_t = 1$), history of CHD 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 CHD and stroke but no adverse event in the current year ($h_t = 4$), survival of a CHD event ($h_t = 5$), survival of a stroke event ($h_t = 6$), death from a non-ASCVD related cause ($h_t = 7$), death from a CHD event ($h_t = 8$), death from stroke ($h_t = 9$), and dead ($h_t = 10$). We use $s_t(d_t, c_t, h_t)$ to denote the state of a patient when we need to specify a component of the state of the patient. Otherwise, we simply use s_t to denote the state of the patient.
- $\mathcal{A}$: action space composed of 0 to 5 antihypertensive medications of five different drug types at their standard dose. Among the types of antihypertensive medications, we include the following: thiazide diuretics (THs), beta-blockers (BBs), calcium channel blockers (CCBs), angiotensin-converting enzyme (ACE) inhibitors, and angiotensin II receptor blockers (ARBs). Since the simultaneous use of ACE inhibitors and ARBs is potentially harmful (Whelton et al. 2018), we exclude the combination of these two drug types from $\mathcal{A}$. Our action space contains a total of 196 treatment choices. The estimates of the effects of antihypertensive drugs on ASCVD events are derived from Law et al. (2009).
- $p_t(s_{t+1}|s_t, a_t)$: transition probability derived from patients' risk for ASCVD events (Goff et al. 2014), the benefit from treatment (Law et al. 2009), fatality likelihoods (Kochanek et al. 2019), and non-ASCVD mortality (Arias and Xu 2019). Based on communications with clinical collaborators, we assume independence among CHD and stroke events. CHD events account for 70% of the ASCVD risk and stroke events account for the remaining 30% (Virani et al. 2020). To be consistent with previous studies, we assume that patients are more likely to have additional CHD or stroke events if they have a history of such ASCVD events (Schell et al. 2016). We account for this by adjusting patients' CHD and stroke odds if they have a history of either ASCVD event (Brønnnum-Hansen et al. 2001, Burn et al. 1994).

- $r_t(s_t, a_t)$: patients' reward given by the quality of life (QoL) weight associated with health condition h_t minus the treatment-related disutility from an antihypertensive medication a_t. The QoL weights and treatment-related disutilities are obtained from previous studies (Kohli-Lynch et al. 2019, Law et al. 2009). We assume that the terminal rewards $r_T(s_T)$ can be computed as the product of the patient's 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 QoL weight (Kohli-Lynch et al. 2019).
- $\alpha(s_t)$: initial state distribution used to represent patients' health condition. Recall that the index t is incorporated into the state definition and represents the effect of patients' age. We select the initial state distribution based on patients' characteristics and test our assumptions in sensitivity analyses.
- γ: discount factor of the model. We use $\gamma = 0.97$ as per recommendations in the medical literature (Neumann et al. 2016).

We use data from the National Health and Nutrition Examination Survey (NHANES) to parameterize our models. Our population is composed of adult Caucasian or African-American patients from 40 to 60 years old with no history of ASCVD. This inclusion criteria leads to a total population of 66.50 million people. To estimate the progression of patients' risk factors over the planning horizon, we linearly regress SBP, DBP, high-density lipoprotein, and total cholesterol on age, age squared, gender, race, smoking status, and diabetes status.

3.3. State and Action Ordering

We order the states of each patient based on their associated risk for ASCVD events. Given the demographic information and clinical observations of a patient, their health condition h_t determines the ordering of the states. Excluding health conditions associated with death, we order patients' states according to the severity of their health condition. This leads to the following order of the states: $s_t(d_t, c_t, 1), s_t(d_t, c_t, 2), s_t(d_t, c_t, 5), s_t(d_t, c_t, 3), s_t(d_t, c_t, 6)$, and $s_t(d_t, c_t, 4)$.

The state classes are also made based on h_t. Given patients' demographic information and clinical observations, we define the following state classes: $\mathcal{S}_1 = \{s_t(d_t, c_t, 1)\}$, $\mathcal{S}_2 = \{s_t(d_t, c_t, 2), s_t(d_t, c_t, 5)\}$,

$\mathcal{S}_3 = \{s_t(d_t, c_t, 3), s_t(d_t, c_t, 6)\}$, and $\mathcal{S}_4 = \{s_t(d_t, c_t, 4)\}$. The first state class includes the states at which patients are healthy, the second class encompasses the states associated with CHD events, the third class covers the states related to stroke events, and the fourth class comprises the states at which patients have a history of both ASCVD events. We did not consider the states with health conditions associated with death, as no treatment is possible in these states.

Actions are ordered as per their effect to the risk for ASCVD events, ranging from no treatment to five standard doses of antihypertensive medications. This order is equivalent to sorting medications according to their expected SBP reductions. Among the same number of antihypertensive doses, we order actions by the estimated risk reduction of each drug type as described in Law et al. (2009). The risk reduction associated with each drug type leads to the following order (from lowest to highest estimated risk reduction): ACE inhibitors, CCBs, THs, BBs, and ARBs. In clinical practice, the drug type selection is often done for patient-specific reasons related to side effects, such as if a patient does not tolerate blood draws or is strongly opposed to leg swelling. But since the difference between the drugs is small, this distinction is likely practically negligible.

We create action classes on the basis of the number of antihypertensive medications being prescribed. The first action class $\mathcal{A}_0$ encompasses the no treatment action and action class $\mathcal{A}_i$ comprises any combination of i antihypertensive medications at standard dose, for $i = 1, ..., 5$. Note that our initial selection of Θ and Ψ satisfy Assumption 1. We study the impact of the state and action classes in our sensitivity analysis.

3.4. Analysis

To understand the implications of interpretable treatment plans at a patient level, we examine the effect of patients' characteristics on π^{CM} and π^{M}. We then study the trade-off between optimality and interpretability at a population level by comparing our policies to π^{*} and π^{G}. We begin by inspecting the number and type of medications recommended by each treatment strategy. Subsequently, we assess the QALYs saved and ASCVD events prevented by each policy, compared to no treatment. Lastly, we inspect the PI for π^{M}, π^{CM}, and π^{G}. To compare if PIs among the policies

are statistically different, we use Wilcoxon Signed Rank Tests with a significance level of 0.05. The significance level and confidence interval (CI) of individual statistical tests are adjusted with the Bonferroni correction method when multiple statistical tests are performed simultaneously.

We study the policy implications of each treatment strategy by dividing our population into BP categories. These categories are created based on the 2017 Hypertension Clinical Practice Guidelines: normal BP, elevated BP, stage 1 hypertension, and stage 2 hypertension. To acknowledge that patients in the NHANES dataset have no history of ASCVD events and the effect of time on their health progression, we assign 99% of the initial state distribution to the states associated with healthy conditions at the first year of our study (i.e., $\alpha(s_0(d_0, c_0, 1)) = 0.99$). The remaining 1% of the initial state distribution is uniformly dispersed over the rest of the states and years.

As the clinical guidelines only provide suggestions regarding the number of antihypertensive medications, we formulate a LP model to find the drug type that maximizes each patient's QALYs. This optimization model follows formulation (1) with additional constraints to guarantee that the number of medications match the recommendations by the clinical guidelines. We limit the total time each LP and MIP spends obtaining an optimal solution to 30 minutes per patient. Any patient exceeding this time limit in any optimization model is excluded from our analysis.

We also perform sensitivity analysis on the treatment strategies by varying our modeling assumptions. Our sensitivity analysis scenarios are described on §EC.4.3 of the e-companion. These scenarios are selected based on communications with our clinical collaborators and information available on the NHANES dataset. In each scenario, we evaluate the PI associated with π^{CM} and π^{M}, the number of ASCVD events allowed due to an interpretable treatment strategy (i.e., the difference between the number of events patients experience following an interpretable policy and π^*), and the average number of medications recommended by each interpretable policy.

3.5. Numerical Results

In this subsection, we examine and describe the effect of the interpretable hypertension treatment plans. We provide insights into the patient- and population-level results in §3.7.

3.5.1. Patient-Level Results. We now evaluate π^{CM} and π^{M} in a series of patient profiles. For comparison purposes, we also determine π^{*} and π^{G} for each patient profile. We first obtain treatment plans for the following patient profile: a 45-year-old, non-diabetic, non-smoker individual with normal BP and normal cholesterol levels. This patient profile will be referred to as the base patient profile. Note that this patient profile does not have any major clinical risk factors for ASCVD. We modify the BP levels of the patient and examine how the policies change.

Figure 1 shows π^{CM} and π^{M} as well as π^{*} over the health conditions of our selection of patient profiles at the last year of our study. The strategies are less intense in earlier years because of our monotonicity restrictions on the actions over time. In the base patient profile, all strategies coincide in recommending no treatment (NT). Thus, there is no PI associated with this profile.

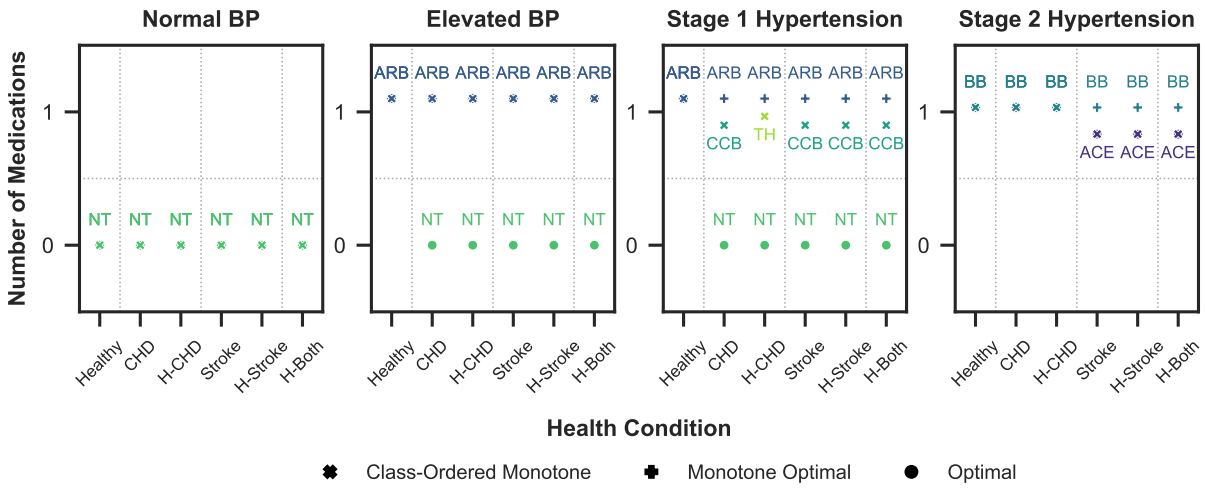

Figure 1 **Treatment policies over the health conditions of selected patient profiles at the last year of our study. The area below and above the horizontal dotted lines represent action classes $\mathcal{A}_0$ and $\mathcal{A}_1$, respectively. The state classes $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$, and $\mathcal{S}_4$ are separated by vertical dotted lines. The label "H-" denotes the state classes associated with a history of ASCVD events. NT: No treatment; ACE: ACE inhibitor; ARB: Angiotensin II receptor blockers; BB: Beta blocker; CCB: Calcium channel blocker; TH: Thiazide.**

Increasing the base profile's arterial pressure level to elevated BP or stage 1 hypertension leads to the suggestion of one ARB at standard dose when the patient has no history of ASCVD in all policies. The optimal policy π^{*} decreases the intensity to no treatment if the patient profile's state

is associated with the survival or history of an ASCVD event. Although optimal, this strategy is not intuitive for physicians or their patients. In contrast, π^M accounts for interpretability aspects by recommending one ARB across all states. Similarly, π^{CM} prescribes one ARB across all states and one CCB or TH in the states associated with ASCVD events for the patient profiles with elevated BP and stage 1 hypertension, respectively. For these patient profiles, there is no considerable consequence for providing interpretability as the PI is less than 0.0001 QALYs. The PI of π^M is larger than the PI of π^{CM} for the profile with stage 1 hypertension, although rather negligible.

When the patient profile has stage 2 hypertension, all the strategies recommend a BB at standard dose if the patient has no history for ASCVD events or has ever survived a CHD event. If the patient has ever survived a stroke, π^{CM} and π^* suggest prescribing one ACE inhibitor. Conversely, since ACE inhibitors have a smaller expected risk reduction than BBs, π^M continues to recommend one BB. While there is no PI associated with π^{CM}, the PI of π^M is positive for this patient profile.

Figure 1 excludes π^G to ease the comparison between π^*, π^{CM}, and π^M. The clinical guidelines π^G recommend NT for normal and elevated BP levels. The profiles with stage 1 and stage 2 hypertension are prescribed one ARB throughout all states, except for the states associated with a history of both ASCVD events. For these states, π^G recommends one ACE inhibitor.

3.5.2. Population-Level Results. We now study the policy implications of adding interpretability restrictions to optimal hypertension treatment plans. Out of a population of 66.50 million people, 27.35 million (41.13%) have normal BP, 12.49 million (18.78%) have elevated BP, 16.51 million (24.83%) have stage 1 hypertension, and 10.15 million (15.26%) have stage 2 hypertension. These findings are consistent with the most recent age-adjusted hypertension prevalence trends across adults in the US (Virani et al. 2020). All the results in this section correspond to patients in the first year of our study.

Treatment Recommendations. The distribution of treatment recommended by each policy per BP category at years 0 and 9 of our study is shown in Figure EC.2 of §EC.4.4 of the e-companion. Other than more intense treatment over time, we note that the distribution of treatment did not change considerably in years 1 through 8.

From the distribution of treatment recommendations, we observe that virtually no patient receives treatment in the normal BP category at any given year. Comparing our interpretable policies to π^* and π^G, we notice that π^{CM} and π^M are often close to optimal. We find that π^* is considerably more aggressive than our interpretable policies at the first year of our study for patients with stage 2 hypertension. Nevertheless, this difference reduces over time to essentially no difference at the last year of our study. We also note that our interpretable policies are typically more intense than π^G for patients with elevated BP and stage 1 hypertension. For patients with stage 2 hypertension, π^G mostly prescribes two to three medications, whereas our interpretable policies fluctuate more broadly from one to three medications.

In terms of the type of medications prescribed, the treatment strategies behave similarly from one to three medications at standard dose. For example, the most frequent medication at standard dose is one ARB across all treatment strategies over time. Two doses of an ARB are prescribed more commonly than two doses of any other drug type or two-drug combination. Similarly, an ARB at three times the standard dose is prescribed more often than any other three-drug combination or three doses. The variation among the drug combinations recommended by the treatment strategies is substantially higher when four medications are prescribed. However, the suggestions from π^{CM} are often close to the recommendations from π^*. For instance, the most common four-drug combinations by both strategies are one ARB at standard dose and three doses of a CCB at year 0 and two doses of an ARB, one BB, plus one CCB at year 9 of our study. Five doses of an ARB are prescribed more regularly than five doses of any other drug type or five-drug combination.

Health Outcomes. As hardly any patient receives treatment under any of the policies in the normal BP category, we focus on patients with elevated BP, stage 1 hypertension, and stage 2 hypertension. We now evaluate the outcomes of patients under each treatment strategy in terms of the number of QALYs saved and ASCVD events prevented, compared to no treatment. In total, π^{CM} and π^M save 12,798 and 12,786 QALYs per 100,000 patients over the planning horizon, compared to no treatment. On the other hand, π^* and π^G save 12,813 and 9,577 QALYs per 100,000

patients, respectively. We notice a similar pattern when comparing the policies in terms of ASCVD events averted. Over the 10-year planning horizon, π^{CM} and π^{M} prevent 1,196 and 1,195 ASCVD events per 100,000 patients, compared to no treatment. The number of ASCVD events prevented by π^{*} and π^{G} are 1,197 and 895 ASCVD events per 100,000 patients, respectively.

Evaluating our results by BP category, we find that patients with stage 2 hypertension receive the greatest benefit from treatment (see Figure EC.3 in §EC.4.4 of the e-companion). We note that patients' health outcomes under π^{CM} and π^{M} are not substantially different for patients with elevated BP or stage 1 hypertension. In people with stage 2 hypertension, π^{CM} saves 79 QALYs and averts 4 ASCVD events more than π^{M} per 100,000 patients. Compared to π^{*}, π^{CM} results in the greatest QALYs loss (33 QALYs per 100,000 patients) and allows the biggest number of ASCVD events (3 events per 100,000 patients) in the elevated BP category. Conversely, π^{M} leads to the greatest QALYs loss (110 QALYs per 100,000 patients) and ASCVD events allowed (6 events per 100,000 patients) in patients with stage 2 hypertension. The clinical guidelines π^{G} are outperformed by our treatment strategies in every BP category. Our policies provide the greatest benefit to patients with elevated BP and stage 1 hypertension, when compared to π^{G}.

Price of Interpretability. Overall, the PI of π^{CM}, π^{M}, and π^{G} are 18, 30, and 3,210 QALYs, respectively. These results are an immediate consequence of the difference in the total QALYs saved between each treatment strategy and π^{*}. In Figure 2, we illustrate the PI corresponding to π^{CM} and π^{M} per BP category, with the normal BP category and outliers above the 99th percentile of the PI in each BP category excluded for illustration purposes. We also show the PI associated with every patient in our dataset following π^{CM} and π^{M} in Figure EC.4 of §EC.4.4.1 of the e-companion.

The PI for π^{CM} and π^{M}, as well as the difference between the two, generally increase with patients' BP. Using Wilcoxon Signed Rank Tests, there is enough evidence to conclude that the PI of π^{M} is significantly greater than the PI of π^{CM} across all patients (95% CI $[0.0001, \infty], P < 0.0001$) and in patients with stage 2 hypertension (98% CI $[0.002, \infty], P < 0.0001$). There was not enough evidence to conclude that PI of π^{CM} is significantly lower than the PI of π^{M} in patients with

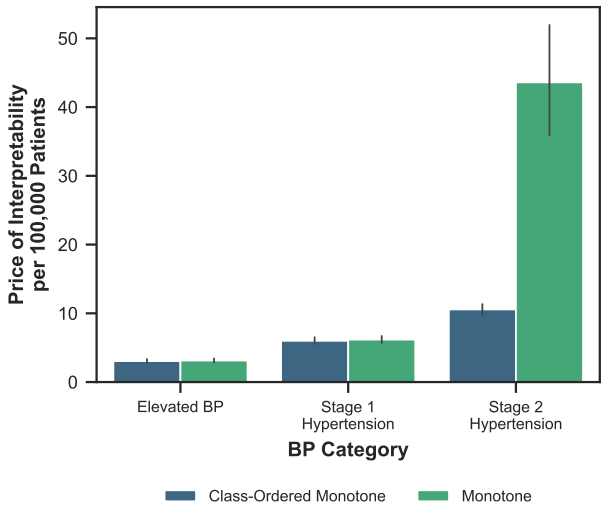

Figure 2 **PI associated with our interpretable policies per BP category. Error bars represent the 95% bootstrap confidence intervals around the PI per 100,000 patients using 10,000 replications.**

elevated BP or stage 1 hypertension. Since we are comparing three BP categories simultaneously, statistical significance was determined using a Bonferroni threshold of 0.02.

The PI of π^G is 4,897, 6,800, and 2,035 QALYs per 100,000 patients with elevated BP, stage 1 hypertension, and stage 2 hypertension, respectively (not shown in Figure 2). Similar to our findings in terms of QALYs saved and ASCVD events prevented, we observe that π^* provides the greatest benefit over π^G for patients with elevated BP and stage 1 hypertension.

3.6. Sensitivity Analyses

We proceed to study how the treatment strategies are affected by changing our modeling assumptions. The results of our sensitivity analyses are summarized in Table 2. Note that the PI and number of ASCVD events allowed in the base case are different than in our main analysis. This difference is due to a larger number of patients exceeding the time limit in the optimization models (30 minutes). In our main analysis, we are excluding 600.00 thousand patients due to the time limit, while in our sensitivity analyses we are excluding 20.50 million patients (see §EC.4.4.2 in the e-companion for details). This exclusion allows us to compare the performance of the policies across the sensitivity analysis scenarios.

Table 2 Sensitivity analyses summary. All results are presented as the average per 100,000 patients. Scenario number matches the enumeration in §EC.4.3. Scenario 0 represents our base case.

Scenario Number	PI		ASCVD Events Allowed[a]		Number of Medications[b]	
	CMP	MP	CMP	MP	CMP	MP
0	21.54	21.56	2.62	2.64	3.61, 16.13, 27.38	3.61, 16.09, 27.38
1	20.68	20.74	2.83	2.83	3.61, 16.13, 27.38	3.61, 16.09, 27.38
2	21.53	21.56	2.62	2.64	3.61, 16.13, 27.38	3.61, 16.09, 27.38
3(a)	21.54	21.56	2.61	2.64	3.61, 16.13, 27.38	3.61, 16.09, 27.38
3(b)	21.95	51.91	2.64	5.37	3.75, 16.39, 27.42	3.49, 15.79, 27.83
4(a)	56.21	59.12	21.01	20.83	2.69, 9.10, 15.57	2.71, 9.08, 15.88
4(b)	47.06	49.40	21.02	21.02	2.82, 8.93, 15.42	2.97, 9.04, 15.24

[a] Results correspond to the first year of our study. Values presented in thousands.

[b] Values represent year 0, year 4, and year 9 of our study, respectively, in thousands.

Ordering the states based on nonincreasing severity (scenario 1), merging the ASCVD events classes into a single class (scenario 2), or categorizing the actions according to their SBP reductions (scenario 3(a)) do not have substantial effects on the outcomes of the policies. Ordering the states on the basis of nonincreasing severity of the health conditions allows the interpretable policies to mimic π^* more closely, which results in marginal reductions in the PI. However, this ordering of the states leads both π^{CM} and monotone policies to prevent fewer ASCVD events, compared to π^*. Combining the ASCVD events classes into a single class offers increased flexibility to π^{CM}, which results in modest reductions in the PI and the ASCVD events allowed. Grouping the actions based on their SBP reductions results in a small decrease in the number of ASCVD events allowed by π^{CM}. None of these scenarios dramatically change the number of medications prescribed over time.

Ordering and categorizing the actions on the basis of their DBP reductions (scenario 3(b)) change the outcomes of the interpretable policies noticeably. Even though π^M treats patients more aggressively, we notice that the PI and ASCVD events allowed are markedly higher in this scenario. Ordering the actions in line with their DBP reductions greatly limits the efficacy of π^M. We also

note that this action ordering and classification leads to worse health outcomes if patients follow π^{CM}, in spite of more aggressive treatment. However, the effect of ordering and classifying the actions based on their DBP reductions is much smaller on π^{CM} than on π^{M}.

Changing the initial state distribution (scenarios 4(a) and 4(b)) also has sizable consequences on the interpretable policies. Recall that π^{CM} and π^{M} generally depend on the initial state distribution (see Proposition 1 and Remark EC.1). We find that the PI and number of ASCVD events allowed by both treatment strategies are higher in the scenarios with modified initial state distribution weights. In these scenarios, the PI is considering the rewards associated with ASCVD events, besides the rewards related to the healthy condition. The effect of suboptimal treatment (due to interpretability constraints) is amplified when these rewards are considered. Furthermore, in these scenarios our interpretable policies are more conservative than in the base case. Lower treatment intensity results in considerable increases in the number of ASCVD events allowed.

3.7. Discussion and Implications

In this subsection, we provide potential explanations for our findings and discuss their implications.

3.7.1. Patient-level Implications. Two key observations can be made at a patient level. First, we notice that π^{*} tends to suggest less treatment as the severity of the health condition increases. This conduct does not reflect physicians' intuition in practice. A potential explanation for this behavior is that the policy aims to maximize the expected discounted QALYs and not to minimize the total number of ASCVD events. As a result, π^{*} focuses on recommending the most aggressive treatment possible to avoid primary events and maintain patients in the healthy condition. The effect of additional treatment in the transition probabilities and, in turn, the rewards is typically smaller than the treatment-related disutility in the states associated with ASCVD events. Second, π^{M} typically recommends to keep a constant treatment across all health conditions in each year of the planning horizon. Similarly, π^{CM} usually keeps the number of medications constant throughout all health conditions at each year of the planning horizon. Rather than reducing the number of medications, π^{CM} generally prescribes a drug combination with lower risk reductions

and treatment-related disutilities. For example, π^{CM} may prescribe CCBs or THs instead of ARBs, and ACE inhibitors instead of BBs. These patterns align with intuition of physicians in practice. Hence, π^{CM} and π^{M} provide more intuitive strategies than π^* with only a small loss in QALYs. However, as the difference among antihypertensive drugs is small, the choice of drug type may be driven by patient-specific reasons instead of QALYs in practice.

3.7.2. Population-level Implications. Several major trends can be noticed at a population level. First, the difference in the number of medications prescribed by our interpretable policies (i.e., π^{CM} and π^{M}) and π^* decreases over time. A reason for this difference in treatment intensity over time is that our interpretable policies are confined to prescribing treatment with nondecreasing intensity over time. Second, all policies tend to include ARBs in their treatment recommendations. This may happen because ARBs have the largest expected ASCVD risk reductions with relatively low treatment-related disutilities. Third, the PI of π^{CM} and π^{M} generally increases with patients' BP. An explanation for this behavior is that as the BP of patients increase, so does the number of medications recommended by the policies. The optimal policy π^* may suggest to decrease the aggressiveness in treatment, whereas the π^{CM} and π^{M} would never decrease treatment intensity. As the number of medication increases, π^* may recommend larger decreases in treatment aggressiveness. Thus, the monotonicity constraints become more restrictive. The pairwise differences between the PI of π^{CM} and π^{M} tend to grow as patients' BP increases for a similar reason. The constraints from the π^{M} will always be at least as restrictive as the restraints from π^{CM}. Fourth, the restrictiveness of π^{CM} normally depends on patients' BP level. Patients with higher BP readings generally receive more treatment. As the number of medications increases, so does the number of potential drug type combinations. A greater number of medications and drug combinations results in larger action classes, which leads to less restrictions in π^{CM}. Finally, we find that π^{CM} offers intuitive treatment strategies to physicians with modest improvements over π^{M}. The optimal CMP π^{CM} results in similar health outcomes to π^* with the added benefits of interpretable recommendations.

3.7.3. Consequences of Changing Modeling Assumptions. In our sensitivity analyses, we find that modifying our modeling assumptions can affect our results in different magnitudes. To a small extent, the ordering and classification of the states can alter the PI and number of ASCVD events averted by our interpretable policies. For example, ordering the states in nonincreasing severity of health condition forces π^{CM} and π^{M} to prescribe at most as much medication in the states associated with ASCVD events as they are in the states related to the healthy condition. Recommending less medication to patients with a history of ASCVD events may prompt additional events. Ordering the actions according to their DBP reductions has a larger impact on the health outcomes associated to our interpretable policies. This finding may be because patients' DBP is not necessary to calculate their risk for ASCVD events. The risk only considers patients' SBP, which implies that the transition probabilities and rewards do not consider patients' DBP. Even if an action is expected to have a high DBP reduction, it may not lead to a high risk reduction. Changing the classification of the actions may have moderate to large effects. For instance, grouping the actions based on their SBP reductions causes the π^{CM} to prevent additional ASCVD events. This decrease may be happening because the action classes according to the SBP reductions are less restrictive than the action classes in line with the number of medications. Conversely, categorizing actions according to their DBP reductions may restrict π^{CM} to classes with more intense treatment that may not lead to larger risk reductions. As a consequence, patients may experience worse health outcomes if the action classes are created according to DBP reductions than if they are defined based on the number of medications or SBP reductions.

4. Conclusions

MDPs are a powerful tool for optimizing and analyzing sequential decisions under uncertainty. Yet, their resulting optimal policy recommendations may follow a structure or pattern that human DMs cannot easily interpret or explain. Due to the lack of structure, there may be a reluctance to implement these policies in practice. To address this issue, we analyzed the PI for the optimal monotone policy and a newly proposed class of structured policies: the CMP. The optimal CMP

may be better aligned with DMs' intuition than the optimal policy while maintaining a lower PI than the optimal monotone policy.

In our case study, we studied the implications of interpretable hypertension treatment plans at a patient and a population level. Our treatment strategies outperformed the current clinical guidelines across all BP categories. This is an indication that the clinical guidelines may be under-treating some patients and over-treating other patients. Consequently, we observed the clinical guidelines resulted in QALY losses compared to our interpretable policies. A reason for this finding may be that our treatment strategies are informed by risk, while the clinical guidelines are mainly driven by BP levels. In addition, the current guidelines do not explicitly consider the patient's future health status and which subsequent clinical conditions are likely to be observed in the future. In contrast, our MDP-derived treatment policies do. These differences may explain why our policies are more aggressive than the clinical guidelines for patients with elevated BP and stage 1 hypertension. Our interpretable policies matched clinicians' intuition with moderate negative consequences in a large population of adults in the US.

The clinical component of this research could be extended by incorporating other conditions, such as high cholesterol or diabetes. Based on communications with our clinical collaborators, we decided to develop interpretable treatment strategies for the management of hypertension as a starting point. Integrating the treatment of multiple conditions will likely increase benefits from our CMP. Our results provide a lower bound on the advantages of the optimal CMP for the management of ASCVD. An alternative to the modeling approach presented in this paper could be to design the state classes based on the factors that influence a patient's ASCVD risk (e.g., SBP, total cholesterol, high-density lipoprotein, and low-density lipoprotein). This alternative approach could result in a larger number of state classes, which may be beneficial to obtain interpretable treatment strategies. However, these state classes would depend on thresholds for each factor, which can be highly subjective. Lastly, measurement error could limit the accuracy of our policies. One crucial form of error is the impact of race and sex on clinical outcomes. Race and gender biases

in the calculation of the risk for ASCVD events may alter cardiovascular outcomes, which could propagate to our treatment recommendations. This vital problem is out of the scope of this work and merits follow-up dedicated to addressing it specifically.

Our work also suggests several future research directions on the technical aspects of designing interpretable policies for MDPs. First, we only considered two classes of interpretable policies: monotone and class-ordered monotone. Future work may propose other interpretable policies for MDP and analyze their respective PIs. Second, we formulated MIPs which can exactly determine the optimal monotone policy and the optimal CMP. However, these solution approaches may be computationally prohibitive when the problem size is large. Future research can investigate computationally efficient algorithms for exactly or approximately obtaining the optimal structured policies. Finally, the MDP considered in this paper consists of state and action spaces which are discrete and finite. Future work can extend our results to investigate the PI for structured policies in more complex state spaces.

In summary, our work provides a foundation for designing policies for MDPs that retain the interpretability of monotone policies with minimal loss in performance compared to the optimal policy. We demonstrate that in complex environments such as personalized hypertension treatment planning, our novel CMPs can make significant improvements over existing guidelines while recommending policies that are aligned with clinicians' intuition. As such, CMPs have great potential to facilitate the implementation of MDP-guided recommendations into practice, with applications in medical decision-making and beyond.

References

Agnihothri S, Cui L, Delasay M, Rajan B (2018) The value of mHealth for managing chronic conditions. *Health Care Management Science* .

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

Ayer T, Zhang C, Bonifonte A, Spaulding AC, Chhatwal J (2019) Prioritizing hepatitis C treatment in U.S. Prisons. *Operations Research* 67(3):853–873.

Bellman R, Glicksberg I, Gross O (1955) On the optimal inventory equation. *Management Science* 2(1):83–104.

Bhattacharya A, Kharoufeh JP (2017) Linear programming formulation for non-stationary, finite-horizon Markov decision process models. *Operations Research Letters* 45(6):570–574.

Boloori A, Saghafian S, Chakkera HA, Cook CB (2020) Data-Driven Management of Post-transplant Medications: An Ambiguous Partially Observable Markov Decision Process Approach. *Manufacturing & Service Operations Management* (February):msom.2019.0797.

Brønnnum-Hansen H, Jørgensen T, Davidsen M, Madsen M, Osler M, Gerdes LU, Schroll M (2001) Survival and cause of death after myocardial infarction: the Danish MONICA study. *Journal of Clinical Epidemiology* 54(12):1244–1250.

Burn J, Dennis M, Bamford J, Sandercock P, Wade D, Warlow C (1994) Long-term risk of recurrent stroke after a first-ever stroke. The Oxfordshire Community Stroke Project. *Stroke* 25(2):333–7.

Capan M, Khojandi A, Denton BT, Williams KD, Ayer T, Chhatwal J, Kurt M, et al. (2017) From data to improved decisions: operations research in healthcare delivery. *Medical Decision Making* 37(8):849–859.

Cevik M, Ayer T, Alagoz O, Sprague BL (2018) Analysis of Mammography Screening Policies under Resource Constraints. *Production and Operations Management* 27(5):949–972.

Chanchaichujit J, Tan A, Meng F, Eaimkhong S (2019) *Optimization, Simulation and Predictive Analytics in Healthcare*, 95–121 (Singapore: Springer Singapore), ISBN 978-981-13-8114-0.

Chen Q, Ayer T, Chhatwal J (2018) Optimal M -Switch Surveillance Policies for Liver Cancer in a Hepatitis C–Infected Population. *Operations Research* 66(3):673–696.

Chobanian AV, Bakris GL, Black HR, Cushman WC, Green LA, Izzo JL, Jones DW, et al. (2003) Seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. *Hypertension* 42(6):1206–1252.

Cohen JB, Townsend RR (2018) The ACC/AHA 2017 Hypertension Guidelines: Both Too Much and Not Enough of a Good Thing? *Annals of Internal Medicine* 168(4):287.

Denton BT, Alagoz O, Holder A, Lee EK (2011) Medical decision making: open research challenges. URL `http://dx.doi.org/10.1080/19488300.2011.619157`.

Denton BT, Kurt M, Shah ND, Bryant SC, Smith Sa (2009) Optimizing the start time of statin therapy for patients with diabetes. *Medical Decision Making* 29(3):351–367.

Goff DC, Lloyd-Jones DM, Bennett G, Coady S, D'agostino RB, Gibbons R, Greenland P, et al. (2014) 2013 acc/aha guideline on the assessment of cardiovascular risk: a report of the american college of cardiology/american heart association task force on practice guidelines. *Journal of the American College of Cardiology* 63(25 Part B):2935–2959.

Hicklin K, Ivy JS, Payton FC, Viswanathan M, Myerse E (2018) Exploring the value of waiting during labor. *Service Science* 10(3):334–353.

Ioannidis JP (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.

James PA, Oparil S, Carter BL, Cushman WC, Dennison-Himmelfarb C, Handler J, Lackland DT, et al. (2014) 2014 Evidence-Based Guideline for the Management of High Blood Pressure in Adults. *Journal of the American Medical Association* 311(5):507.

Kochanek KD, Murphy SL, Xu J, Arias E (2019) Deaths: final data for 2017. *National Vital Statistics Reports* 68(9):1–18.

Kohli-Lynch CN, Bellows BK, Thanassoulis G, Zhang Y, Pletcher MJ, Vittinghoff E, Pencina MJ, Kazi D, Sniderman AD, Moran AE (2019) Cost-effectiveness of Low-density Lipoprotein Cholesterol Level–Guided Statin Treatment in Patients With Borderline Cardiovascular Risk. *JAMA Cardiology* 4(10):969–977.

Kurt M, Denton BT, Schaefer AJ, Shah ND, Smith Sa (2011) The structure of optimal statin initiation policies for patients with Type 2 diabetes. *IIE Transactions on Healthcare Systems Engineering* 1(July):49–65.

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

Law MR, Morris JK, Wald NJ (2009) Use of blood pressure lowering drugs in the prevention of cardiovascular disease: meta-analysis of 147 randomised trials in the context of expectations from prospective epidemiological studies. *BMJ* 338:b1665.

Lee E, Lavieri MS, Volk M (2018) Optimal screening for hepatocellular carcinoma: a restless bandit model. *Manufacturing & Service Operations Management* (January 2019):msom.2017.0697.

Littman ML (1994) Memoryless policies: Theoretical limitations and practical results. *From Animals to Animats 3: Proceedings of the third international conference on simulation of adaptive behavior*, volume 3, 238 (Cambridge, MA).

Lovejoy WS (1987) Some monotonicity results for partially observed markov decision processes. *Operations Research* 35(5):736–743.

Marrero WJ, Lavieri MS, Sussman JB (2021) Optimal cholesterol treatment plans and genetic testing strategies for cardiovascular diseases. *Health Care Management Science* .

Mason JE, Denton BT, Shah ND, Smith SA (2014) Optimizing the simultaneous management of blood pressure and cholesterol for Type 2 diabetes patients. *European Journal of Operational Research* 233(3):727–738.

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

Pandya A, Sy S, Cho S, Weinstein MC, Gaziano TA (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.

Petrik M, Luss R (2016) Interpretable policies for dynamic product recommendations. *32nd Conference on Uncertainty in Artificial Intelligence 2016, UAI 2016*, 607–616, ISBN 9781510827806.

Puterman ML (2014) *Markov decision processes: discrete stochastic dynamic programming* (John Wiley & Sons).

Saville CE, Smith HK, Bijak K (2018) Operational research techniques applied throughout cancer care services: a review. *Health Systems* 6965:1–22.

Schäl M (1976) On the optimality of (s,s)-policies in dynamic inventory models with finite horizon. *SIAM Journal on Applied Mathematics* 30(3):528–537.

Schell GJ, Marrero WJ, Lavieri MS, Sussman JB, Hayward RA (2016) Data-driven Markov decision process approximations for personalized hypertension treatment planning. *MDM Policy & Practice* 1(1).

Serfozo RF (1976) Monotone optimal policies for markov decision processes. *Stochastic Systems: Modeling, Identification and Optimization, II*, 202–215 (Springer).

Serin Y, Kulkarni VG (1995) Implementable Policies: Discounted Cost Case. *Computations with Markov Chains* 283–306.

Skandari MR, Shechter SM (2021) Patient-Type Bayes-Adaptive Treatment Plans. *Operations Research* (March):opre.2020.2011.

Smith JE, McCardle KF (2002) Structural properties of stochastic dynamic programs. *Operations Research* 50(5):796–809.

Solberg LI, Miller WL (2018) The new hypertension guideline: logical but unwise. *Family Practice* 35(5):528–530.

Steimle LN, Kaufman DL, Denton BT (2021) Multi-model Markov decision processes. *IISE Transactions* 1–16.

Suen Sc, Brandeau ML, Goldhaber-Fiebert JD (2018) Optimal timing of drug sensitivity testing for patients on first-line tuberculosis treatment. *Health Care Management Science* 21(4):632–646.

Van Der Wardt V, Harrison JK, Welsh T, Conroy S, Gladman J (2017) Withdrawal of antihypertensive medication: A systematic review. *Journal of Hypertension* 35(9):1742–1749.

Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, et al. (2020) *Heart disease and stroke statistics—2020 update: A report from the American Heart Association.*

Vlassis N, Littman ML, Barber D (2012) On the computational complexity of stochastic controller optimization in pomdps. *ACM Transactions on Computation Theory (TOCT)* 4(4):1–8.

Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Dennison Himmelfarb C, DePalma SM, 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.

Williams B, Mancia G, Spiering W, Agabiti Rosei E, Azizi M, Burnier M, Clement DL, et al. (2018) 2018 ESC/ESH guidelines for the management of arterial hypertension. *European Heart Journal* 39(33):3021–3104.

E-Companion
An Analysis of Structured Optimal Policies for Hypertension Treatment Planning: The Tradeoff Between Optimality and Interpretability

EC.1. Proofs of Technical Results

[Proposition 1] The optimal monotone policy π^M depends on the initial distribution α.

Proof of Proposition 1 We prove this result by providing an example. Consider the MDP in Figure EC.1 and the corresponding value-to-go matrix described in Table EC.1. Based on this table, the optimal policy π^* is $(1,2,1,\cdot,\cdot)$ where states 4 and 5 can take either action 1 or 2. This policy gives $J^{\pi^*}(\alpha)=\sum_{t=1}^{\infty}\gamma^t$ regardless of the initial distribution α. If the initial distribution is $\alpha^1=(0.5,0,0.5,0,0)$, one optimal monotone policy is $\pi^1=(1,1,1,1,1)$ which gives an expected value of $J^{\pi^1}(\alpha^1)=\sum_{t=1}^{\infty}\gamma^t$. However, if the initial distribution is $\alpha^2=(0,1,0,0,0)$, then $J^{\pi^1}(\alpha^2)=0$, while $\pi^2=(1,2,2,2,2,2)$ is a monotone policy that has value $J^{\pi^2}(\alpha^2)=\sum_{t=1}^{\infty}\gamma^t>0$. $\square$

Remark EC.1. To show that the optimal CMP is also history-dependent, it suffices to consider the same example in the proof of Proposition 1 with state classes $\mathcal{S}_1=\{1,2\}$ and $\mathcal{S}_2=\{3,4,5\}$. In this case, we do not follow the trivial example of setting the state and action classes to follow strict monotonicity, but still end up with the same result.

Table EC.1 The value to go corresponding to each state

State	Action	v(s,a)
1	1	$\sum_{t=1}^{\infty}\gamma^t$
1	2	0
2	1	0
2	2	$\sum_{t=1}^{\infty}\gamma^t$
3	1	$\sum_{t=1}^{\infty}\gamma^t$
3	2	0

[Lemma 1] The number of monotone policies is given by $\binom{S+A-1}{A-1}$.

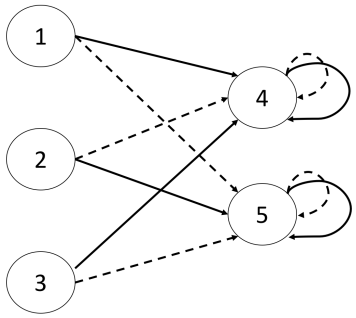

Figure EC.1 **An MDP for which the optimal monotone policy, $\bar{\pi}$, depends on the initial distribution, α. The solid (dashed) lines represent transitions corresponding to taking action 1 (action 2) which occur with probability 1. The rewards are defined as $r(s,a)=0$ for all $s\neq 4, a\in\mathcal{A}$ and $r(4,a)=1$, for all $a\in\mathcal{A}$.**

Proof of Lemma 1. We prove this result by induction for the case of non-decreasing policies. Let $Pol(S,A)$ denote the number of policies for an MDP with S states and A actions. We will establish that $Pol(S,A)=Pol(S-1,A)+Pol(S,A-1)=\binom{S+A-1}{A-1}$.

First, consider a base case with $A=1$ actions and S states. In this case, there is only 1 monotone policy (i.e., $\pi(s)=1,\forall s\in\{1,\ldots,S\}$). Therefore, the base case $Pol(S,1)=\binom{S+1-1}{1-1}=\binom{S}{0}=1$ holds. Now, consider $S=1$ and A actions. There are A possible monotone policies (i.e., $\pi=a$ for each $a\in\mathcal{A}$). Therefore, $Pol(1,A)=\binom{1+A-1}{A-1}=\binom{A}{A-1}=A$ holds.

By induction, we now establish that $Pol(S,A)=Pol(S-1,A)+Pol(S,A-1)$ for any positive integers S,A. Assume that the base case holds for both $Pol(S-1,A)$ and $Pol(S,A-1)$. When considering the number of monotone policies for the MDP with S states and $A-1$ actions, each of these $Pol(S,A-1)$ monotone policies remains a valid monotone policy in the case where there are A actions. Now, consider each of the policies for the MDP where there are $S-1$ and A actions. Each of the monotone policies for MDP remains a valid monotone policy so long as it is appended with $\pi(S)=A$. Furthermore, since each of these policies contain the action A, this set of policies does not intersect with the set of policies for S states and $A-1$ actions. This construction comprises all of the valid monotone policies. Therefore, we have established that

$Pol(S,A) = Pol(S-1,A) + Pol(S,A-1)$. By the induction hypothesis, we have

$$Pol(S,A) = \binom{S+A-2}{A-1} + \binom{S+A-2}{A-2} = \binom{S+A-1}{A-1}, \tag{EC.1}$$

which completes the proof. □

[Corollary 1] The number of monotone policies grows as a polynomial in S and A.

Proof of Corollary 1. The resulting value in Lemma 1 is a binomial coefficient involving the number of states and number of actions. The number of monotone policies for various problem sizes closely relates to the numbers on the diagonals of Pascal's triangle. For a fixed number of states S, the number of monotone policies corresponds to the A^{th} number on the $S+1$ diagonal of Pascal's triangle. For a fixed number of actions A, the number of monotone policies corresponds to the S^{th} number of the A^{th} diagonal of Pascal's triangle. For each diagonal of Pascal's triangle, the numbers on the diagonal grow as a polynomial function and therefore, the number of monotone policies grows as a polynomial in the number of states for a fixed $\mathcal{S}$ and as a polynomial in the number of actions for a fixed $\mathcal{A}$. □

[Lemma 2] Consider any class functions Θ, Ψ which satisfy Assumption 1.

1. Suppose that Θ admits at least two state classes. Let Θ' be a new state class function that merges two classes admitted by Θ, i.e., $\Theta'(s) = \Theta'(s') = k$ for all s, s' where $\Theta(s) = k$ and $\Theta(s') = k+1$ for an arbitrary $k \in \mathcal{K} \setminus \{K\}$. Then $\pi \in \Pi^{CM}_{\Theta,\Psi}$ implies $\pi \in \Pi^{CM}_{\Theta',\Psi}$.
2. Suppose that Ψ admits at least two action classes. Let Ψ' be a new action class function that merges two classes admitted by Ψ. Then, $\pi \in \Pi^{CM}_{\Theta,\Psi}$ implies $\pi \in \Pi^{CM}_{\Theta,\Psi'}$

Proof of Lemma 2. To show that Property 1 holds, it suffices to show that for any s^* in the newly merged class and policy $\pi \in \Pi^{CM}_{\Theta,\Psi}$, having $\Theta'(s^*) \geq \Theta'(s)$ implies $\Psi(\pi(s^*)) \geq \Psi(\pi(s))$ and having $\Theta'(s^*) \leq \Theta'(s)$ implies $\Psi(\pi(s^*)) \leq \Psi(\pi(s))$. Take any s^* in the newly merged class and s such that $\Theta'(s^*) \geq \Theta'(s)$. By the construction of Θ', it follows that $\Theta(s^*) \geq \Theta(s)$. Since $\pi \in \Pi^{CM}_{\Theta,\Psi}$, we have $\Psi(\pi(s^*)) \geq \Psi(\pi(s))$. Showing the reverse inequality is similar. Hence, Property 1 has been shown. The proof for showing Property 2 is similar and has been omitted. □

[Theorem 2] For any arbitrary class functions Θ and Ψ which satisfy Assumption 1, we have

$$\max_{\pi\in\Pi^M} J^\pi(\alpha) \leq \max_{\pi\in\Pi^{CM}_{\Theta,\Psi}} J^\pi(\alpha). \tag{EC.2}$$

Proof of Theorem 2. We specifically show that for any Θ and Ψ resulting in K and G state and action classes, respectively, inequality (EC.2) holds. We show this result by backward induction.

Part 1: We first show the desired result for any state class function Θ. Fix Ψ such that $\Psi(a) = a$ for all $a \in \mathcal{A}$. To show the base case, consider the class function Θ_S, which results in S state classes with each state being its own class. Since Θ satisfies Assumption 1, it must be the case that $\Theta_S(s) = s$ for all $s \in \mathcal{S}$. That is, the class structure imposed by Θ_S is exactly the same as the strict state-wise ordering imposed by standard monotone policies. Hence, we have

$$\max_{\pi\in\Pi^{CM}_{\Theta_S,\Psi}} J^\pi(\alpha) = \max_{\pi\in\Pi^M} J^\pi(\alpha).$$

Hence, the base case holds. Now, assume that for any arbitrary monotone state class function resulting in $S, S-1, ..., k+1$ classes, that inequality (EC.2) holds. Then, it remains to show that the inequality holds for Θ resulting in k classes. Let Θ_k denote an arbitrary state class function resulting in k state classes. Take any state s such that it is either the greatest or least member of its class and separate it so that it becomes its own class. Let Θ_{k+1} denote the class function for this newly constructed class structure with $k+1$ classes. Additionally, let $\pi^* = \arg\max_{\pi\in\Pi^{CM}_{\Theta_{k+1},\Psi}} J^\pi(\alpha)$. From Property 1 of Lemma 2, we have $\pi^* \in \Pi^{CM}_{\Theta_k,\Psi}$. Therefore, by the induction step and the definition of optimality, we have

$$\max_{\pi\in\Pi^{CM}_{\Theta_k,\Psi}} J^\pi(\alpha) \geq J^{\pi^*}(\alpha) = \max_{\pi\in\Pi^{CM}_{\Theta_{k+1},\Psi}} J^\pi(\alpha) \geq \max_{\pi\in\Pi^M} J^\pi(\alpha).$$

Hence, we have shown that for any state class function Θ which satisfies Assumption 1, inequality (EC.2) holds. That is, we can obtain a greater value function than a standard monotone policy through any monotone state class function.

Part 2: We now show that the inequality (EC.2) holds for any arbitrary action class function Ψ. Take any arbitrary state class function Θ and let Ψ_A denote the action class function resulting

in A action classes. Since Ψ_A is monotone, it must be the case that $\Psi(a) = a$ for all $a \in \mathcal{A}$. Hence, the base case holds by Part 1 of this proof. Now, assume that inequality (EC.2) holds for all Ψ resulting in $A, A-1, ..., g+1$ classes. Let Ψ_g denote an arbitrary action class function resulting in g action classes. Now, take any action a which is either the largest or smallest member of its class and separate it so that it becomes its own class. Let Ψ_{g+1} denote the action class function corresponding to this newly constructed action class structure. Additionally, let $\pi^* = \arg\max_{\pi \in \Pi^{CM}_{\Theta,\Psi_{g+1}}} J^\pi(\alpha)$. From Property 2 of Lemma 2, it follows that $\pi^* \in \Pi^{CM}_{\Theta,\Psi_g}$. By the induction step and definition of optimality, we have

$$\max_{\pi \in \Pi^{CM}_{\Theta,\Psi_g}} J^\pi(\alpha) \geq J^{\pi^*}(\alpha) = \max_{\pi \in \Pi^{CM}_{\Theta,\Psi_{g+1}}} J^\pi(\alpha) \geq \max_{\pi \in \Pi^M} J^\pi(\alpha).$$

Hence, we have shown that (EC.2) holds for the given Θ and any action class function Ψ. Furthermore, since the Θ was chosen arbitrarily, we can state more generally that (EC.2) holds for any monotone Θ and Ψ. $\square$

EC.2. Dual Formulations of M-MIP and CM-MIP

The dual formulation of M-MIP is given by

$$\text{(M-MIP-D)} \quad \max_{\mathbf{x},\mathbf{y}} \quad \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s,a) y(s,a) \tag{EC.3a}$$

$$\text{subject to:} \quad \sum_{a \in \mathcal{A}} y(s,a) - \gamma \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} P(s|s',a') y(s',a') = \alpha(s) \text{ for all } s \in \mathcal{S} \tag{EC.3b}$$

$$\sum_{a \in \mathcal{A}} x(s,a) = 1 \text{ for all } s \in \mathcal{S} \tag{EC.3c}$$

$$y(s,a) \leq M x(s,a) \text{ for all } s \in \mathcal{S},\ a \in \mathcal{A} \tag{EC.3d}$$

$$x(s,a) \leq \sum_{a' \geq a} x(s+1, a') \text{ for all } s \in \mathcal{S} \setminus S, a \in \mathcal{A} \tag{EC.3e}$$

$$y(s,a) \geq 0 \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}. \tag{EC.3f}$$

$$x(s,a) \in \{0,1\} \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}. \tag{EC.3g}$$

In (EC.3d), the parameter M is a large constant. It is well-known that the dual variable $y(s,a)$ represents a discounted "count" of being in state s and performing action a, i.e., $y(s,a) =$

$\sum_{t=0}^{\infty}\gamma^t\mathbb{P}(s_t=s,a_t=a)$. Hence, we can set $M=\frac{1}{1-\gamma}=\sum_{t=0}^{\infty}\gamma^t\geq\sum_{t=0}^{\infty}\gamma^t\mathbb{P}(s_t=s,a_t=a)$ as an upper bound. Note that in the finite horizon case, $y_t(s,a)=\mathbb{P}(s_t=s,a_t=a)$ so setting $M=1$ suffices. Furthermore, by modifying M-MIP-D, the dual formulation for CM-MIP is given by

$$\text{(CM-MIP-D)}\quad \max_{\mathbf{x},\mathbf{y}}\quad \text{(EC.3a)}\quad \text{subject to:}\quad \text{(EC.3b)-(EC.3d), (EC.3f)-(EC.3g), (5).}$$

EC.3. Warm Start via Monotone Policy Iteration

Typically, the computational effort required to solve an MIP can be reduced through a *warm start*, i.e., supplying an initial feasible solution. Here, we modify the classic policy iteration algorithm to identify a policy which is guaranteed to be monotone. Given the importance of the initial distribution α, we rely on α to provide an order by which to prioritize states in the policy. Then, we construct a feasible set of actions for each state based on actions chosen in preceding states. Throughout our algorithm, we denote $\alpha_{[i]}$ as the i^{th} order statistic of α. We initialize our algorithm with some policy $\pi_0(s)$ which is associated with some initial value function v_0. The *Monotone Policy Iteration* algorithm is summarized in Procedure 1.

Much like the classic policy iteration algorithm, the *Monotone Policy Iteration* algorithm terminates in a finite number of iterations (since there are a finite number of monotone policies). However, it is not guaranteed to provide the optimal monotone policy. Regardless, it is straightforward to verify that *Monotone Policy Iteration* generates a monotone policy. Hence, it can be used to warm start M-MIP. Furthermore, since monotone policies are also CMPs (under Assumption 1), this policy can also be used to warm start CM-MIP.

Remark EC.2. A monotone policy iteration algorithm is presented in Puterman (2014, Ch. 6.11.2). The main difference between the algorithm presented in Procedure 1 and this previously developed algorithm is the order in which states are visited to obtain the monotone policy. Specifically, our algorithm visits states according to their prominence in the initial state distribution whereas the algorithm described in Puterman (2014) visits the states in ascending order. Nevertheless, both algorithms can be shown to terminate in finitely many iterations and are guaranteed to provide an optimal policy if the optimal policy is indeed monotone in s.

Procedure 1 Monotone Policy Iteration algorithm

Data: α, $\mathbf{v}_0$, π_0, $\epsilon \geq 0$, $t = 1$

Result: Monotone policy π

while $t = 1$ *or* $\|\boldsymbol{v}_t - \boldsymbol{v}_{t-1}\| \leq \epsilon$ **do**

  for $i = 1$ *to* S **do**

    Set s as the state corresponding to $\alpha_{[i]}$.

    Set $\mathcal{S}^- \leftarrow \{s' \in \bar{\mathcal{S}} : s' < s\}$ and $\mathcal{S}^+ \leftarrow \{s' \in \bar{\mathcal{S}} : s' > s\}$.

    Set $A_{\min} \leftarrow \begin{cases} 1 & \text{if } |\mathcal{S}^-| = 0 \\ \pi_t(\max \mathcal{S}^-) & \text{otherwise.} \end{cases}$ and $A_{\max} \leftarrow \begin{cases} A & \text{if } |\mathcal{S}^+| = 0 \\ \pi_t(\min \mathcal{S}^+) & \text{otherwise.} \end{cases}$.

    Set $\pi_t(s) \leftarrow \arg\max_{\mathcal{A}_{\min} \leq a \leq \mathcal{A}_{\max}} r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) v_{t-1}(s')$.

    Set $\bar{\mathcal{S}} \leftarrow \bar{\mathcal{S}} \cup \{s\}$.

    Perform policy evaluation on π_t to obtain $\mathbf{v}_t$. **if** $\|\boldsymbol{v}_t - \boldsymbol{v}_{t-1}\| \leq \epsilon$ **then**

      return $\pi = \pi_t$

    else

      Set $t \leftarrow t + 1$.

    end

  end

end

EC.4. Case Study Details

We provide additional methods and results from our case study in this section of the e-companion.

EC.4.1. Finite-Horizon MDP Formulations

In this subsection, we adapt our formulations to the context finite-horizon MDPs for the management of hypertension. Because there is no evidence that hypertension treatment is beneficial at low BP levels, our models do not allow for treatment if patients' SBP is below 120 mm Hg or their DBP is below 55 mm Hg. In addition, since many believe hypertension is especially dangerous at high levels, we always offer treatment if patients BP is above 150/90 mm Hg (Schell et al. 2016). We incorporate these clinical restrictions by adding a constraint that establishes that any treatment

choice that violates the minimum SBP or DBP levels cannot be optimal. The set of treatment choices that leads to "clinically infeasible actions" at state s_t and year t is denoted by $I_t(s_t) \subset \mathcal{A}$. This set of actions for each state is identified before formulating the models based on the estimated effect of antihypertensive drugs on the risk for ASCVD events (Law et al. 2009).

To account for the effect of age in the risk of ASCVD events, we modify the LP formulation in (1) as described in Theorem 1 of Bhattacharya and Kharoufeh (2017). The possibility of "clinically infeasible actions" is incorporated into this formulation by constraining the value functions using the rewards of actions that are clinically feasible. That is, by repeating constraint (1b) for all $t \in \mathcal{T}' \setminus \{T\}$, $s_t \in \mathcal{S}$, and $a_t \in \mathcal{A} \setminus I_t(s_t)$. This possibility is incorporated into the dual LP formulation by constraining the dual variables $y_t(s_t, a_t)$ associated with the "clinically infeasible actions" as $\sum_{a_t \in I_t(s_t)} y_t(s_t, a_t) = 0$ for all $t \in \mathcal{T}' \setminus \{T\}$ and $s_t \in \mathcal{S}$. Both approaches result in non-binding constraints at any treatment choice that violates the minimum SBP or DBP levels.

The M-MIP formulation in (3) is modified to account for the nonstationarity of the risk for ASCVD events and the plausibility of "clinically infeasible actions" in a similar way. We incorporate a summation over $t \in \mathcal{T}'$ in (3a), repeat constraint (3b) for all $t \in \mathcal{T}' \setminus \{T\}$, $s_t \in \mathcal{S}$, and $a_t \in \mathcal{A}$, constraint the terminal value functions with the terminal rewards (i.e., $v_T(s_t) \leq r_T(s_T)$ for all $s_T \in \mathcal{S}$), replicate (3c) for all $t \in \mathcal{T}' \setminus \{T\}$ and $s_t \in \mathcal{S}$, repeat constraint (3d) for all $t \in \mathcal{T}' \setminus \{T\}$, $s_t \in \mathcal{S} \setminus S_{T-1}$, and $a_t \in \mathcal{A}$, and restrict the binary variables $x_t(s_t, a_t)$ using the following equation:

$$\sum_{a_t \in I_t(s_t)} x_t(s_t, a_t) = 0 \text{ for all } t \in \mathcal{T}' \setminus \{T\}, s_t \in \mathcal{S}. \tag{EC.5}$$

Note that we have added the index t to the binary variables $x_t(s_t, a_t)$ to highlight their dependence on the decision epoch. The dual formulation of M-MIP in (EC.3) of Section EC.2 of the e-companion is modified as follows:

1. Incorporating a summation over $t \in \mathcal{T}' \setminus \{T\}$ and adding a summation of the product of terminal rewards and dual variables (i.e., $\sum_{s_T \in \mathcal{S}} r_T(s_T) y_T(s_T)$) in (EC.3a).
2. Repeating constraints (EC.3b) and (EC.3c) for all $t \in \mathcal{T}' \setminus \{T\}$ and $s_t \in \mathcal{S}$.

3. Replicating (EC.3d) and (EC.3f) for all $t \in \mathcal{T}' \setminus \{T\}$, $s_t \in \mathcal{S}$, and $a_t \in \mathcal{A}$. Notice that it suffices to take $M = 1$ since $y_t(s,a) = \mathbb{P}(s_t = s, a_t = a)$.
4. Repeating constraint (EC.3e) for all $t \in \mathcal{T}' \setminus \{T\}$, $s_t \in \mathcal{S} \setminus S_{T-1}$, and $a_t \in \mathcal{A}$.
5. Restraining the binary variables $x_t(s_t, a_t)$ using constraint (EC.5).

We adjust the CM-MIP formulation in (6) in a similar manner to (3). However, we repeat constraint (5) for all $t \in \mathcal{T}' \setminus \{T\}$, $s_t \in \mathcal{S}_k$, and $s'_t \in \mathcal{S}_{k+1}$ with $k = 1, \ldots, K-1$ instead of (EC.3e) for all $t \in \mathcal{T}' \setminus \{T\}$, $s_t \in \mathcal{S} \setminus S_{T-1}$, and $a_t \in \mathcal{A}$. The dual formulation of the CM-MIP formulation is also similar to the dual formulation of M-MIP, with the exception that (EC.3e) is modified to (5).

EC.4.2. Enforcing Monotonicity on Decision Epochs.

To guarantee nondecreasing actions over time, we incorporate the following constraint to formulations (3) and (EC.3):

$$x_t(s,a) \leq \sum_{a' \geq a} x_{t+1}(s,a') \text{ for all } t \in \mathcal{T}' \setminus \{T\}, s \in \mathcal{S} \setminus \{S\}, a \in \mathcal{A}.$$

We have dropped the index t in the state s and action a to indicate explicitly that they remain constant in the year. We also add the following constraint to formulation (6) and its dual:

$$\sum_{a \in \mathcal{A}_g} x_t(s,a) \leq \sum_{a' \geq \min \mathcal{A}_g} x_{t+1}(s,a') \text{ for all } t \in \mathcal{T}' \setminus \{T\}, s \in \mathcal{S}, g = 1, ..., G.$$

EC.4.3. Description of Sensitivity Analysis Scenarios

We consider the following sensitivity analysis scenarios:

1. *State order*: In this scenario, we order states by nonincreasing risk for ASCVD events (i.e., $s_t(d_t, c_t, 4), s_t(d_t, c_t, 6), s_t(d_t, c_t, 3), s_t(d_t, c_t, 5), s_t(d_t, c_t, 2)$, and $s_t(d_t, c_t, 1)$).
2. *State classes*: We modify the state classes from one class per ASCVD event to one class encompassing all ASCVD events. In this scenario, we combine $\mathcal{S}_2$, $\mathcal{S}_3$, and $\mathcal{S}_4$ into a new class $\mathcal{S}'_2$. Class $\mathcal{S}_1$ remains unchanged.
3. *Action order and classes*: We examine two scenarios by categorizing each treatment in terms of their expected BP reduction.

(a) We generate five classes using 5 mm Hg increments on the expected SBP reduction of each treatment until 25 mm Hg (slightly above the maximum SBP reduction with 5 medications). As ordering actions according to their SBP reductions is equivalent to ordering them by ASCVD risk reduction, the order of actions within each class remains unchanged.

(b) We classify medications into six classes based on 3 mm Hg increments of each treatment's expected DBP reduction until 18 mm Hg (marginally above the maximum DBP reduction with 5 medications). In this scenario, the actions are ordered in terms of their expected DBP reductions as reported by Law et al. (2009).

4. *Initial state distribution*: We evaluate two additional cases.

(a) To represent the influence of time on patients' health, we assign 99% of the state distribution weight uniformly to states at the first year of our study. The remaining 1% of the initial state distribution is uniformly dispersed over the rest of the states and years.

(b) We weigh α uniformly over all states and years.

EC.4.4. Additional Case Study Results

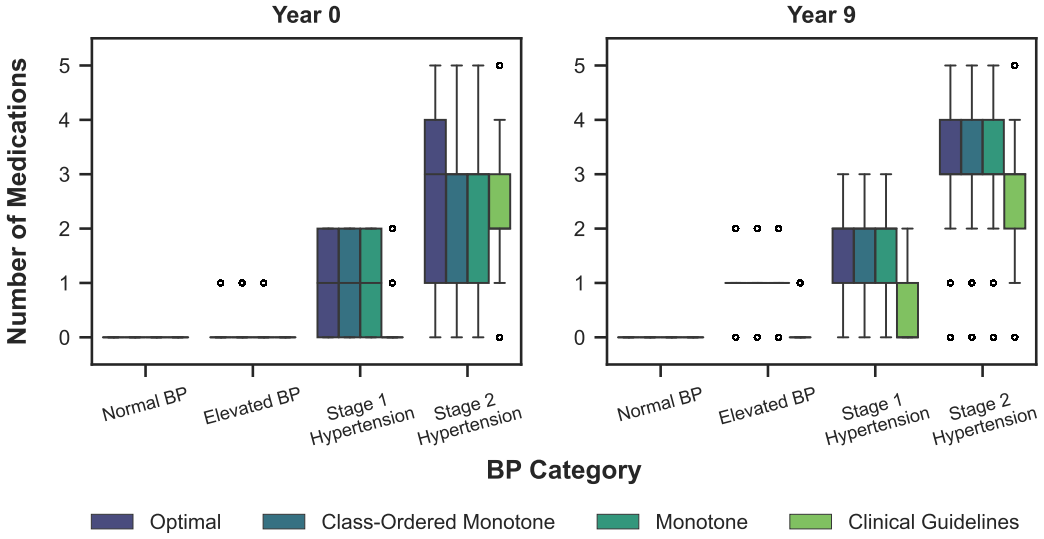

Figure EC.2 **Distribution of treatment at year 0 and year 9 of the study. BP categories are made based on patients' characteristics at year 0.**

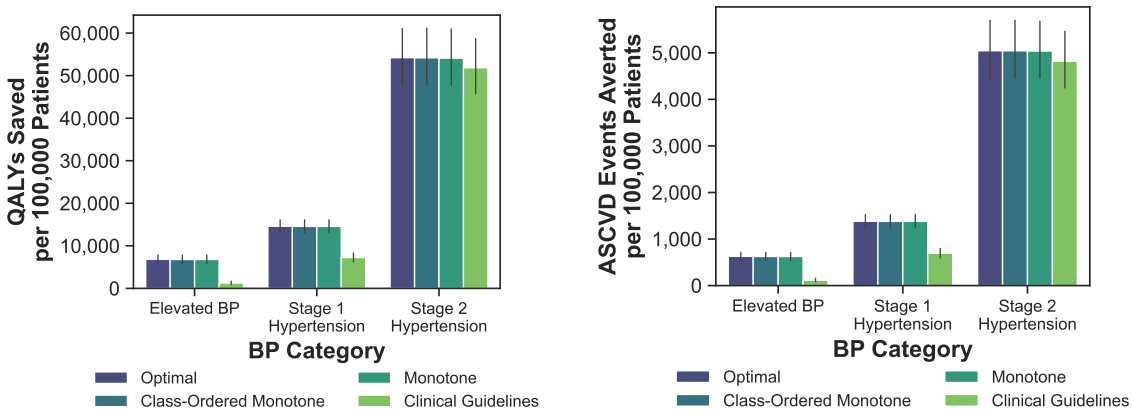

Figure EC.3 **QALYs saved (left) and ASCVD events averted (right) by each treatment policy at every BP category per 100,000 patients, compared to no treatment. Error bars represent the 95% bootstrap confidence intervals around the mean per 100,000 patients using 10,000 replications.**

EC.4.4.1. Distribution of PI. In this subsection, we study the PI of π^{CM} and π^{M} across each patient in our population. For comparison purposes, we also examine the pairwise differences between the total discounted reward of π^{CM} and π^{M} for every patient. We inspect our results as a function of each patient's risk for ASCVD events. The ASCVD risk summarizes the health and provides a rich description of the patient in a single number. In Figure EC.4, we find that the PI of π^{CM} and π^{M} are considerably related. Overall, 2.85% of patients experience a higher PI with π^{M} than with π^{CM}. This translates to 0.92%, 1.36%, and 16.11% for patients with elevated BP, stage 1 hypertension, and stage 2 hypertension. We also note that higher PIs typically appear in lower risk scores at each BP category. Moreover, higher PIs tend to lead to large pairwise differences in PI between the π^{CM} and π^{M}.

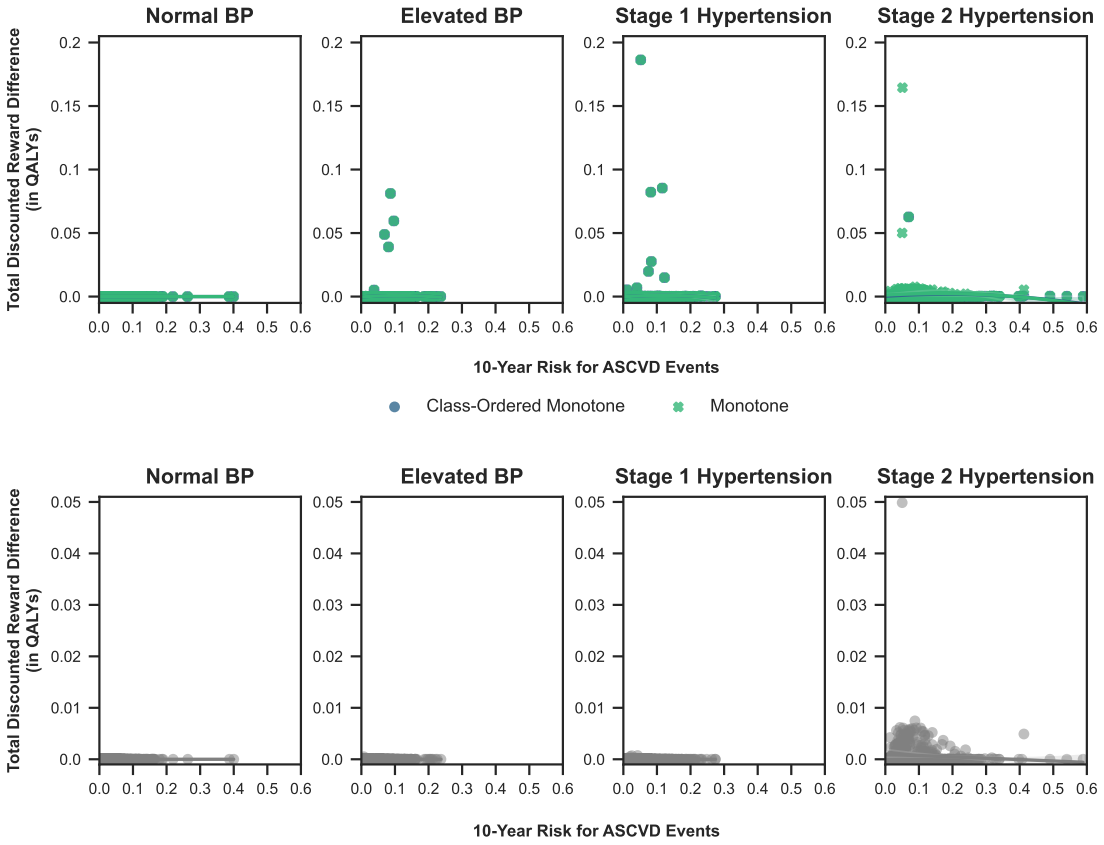

Figure EC.4 **Distribution of PI (top) and pairwise differences between π^{CM} and π^{M} (bottom) over patients' 10-year risk for ASCVD events. Smoothed lines are obtained using second degree local regression. Shaded areas around the smoothed lines represent 95% bootstrapped confidence intervals around the mean using 10,000 replications.**

EC.4.4.2. Effect of Modeling Assumptions on Computational Time. Changing our modeling assumptions affects the computational time of our optimization models, and therefore, the number of patients for whom we obtain optimal solutions. We find that the ordering (scenario 1) and classification (scenario 2) of the states have minor effects on computational time. In these scenarios, no more than 600,000 patients exceed the time limit.

Changing the classification of the actions (scenarios 3(a) and 3(b)) may have moderate to large effects on the time it takes to obtain optimal solutions. A total of 630,000 and 1.4 million patients are excluded in the scenarios where actions are categorized according to their SBP (scenario 3(a))

and DBP (scenario 3(b)) reductions, respectively. A potential explanation for the increase in computational time may be that there is a greater number of action classes in both of these scenarios.

The largest increase in computational time is observed in the scenarios where the initial state distribution is uniformly allocated across multiple states (scenarios 4(a) and 4(b)). We exclude 17.70 million patients in the scenario in which 99% of the initial state distribution is uniformly dispersed across the states representing the first year of our study (scenario 4(a)). A total of 18.93 million patients exceeded the 30-minute time limit when the initial state distribution is uniformly allocated across all states (scenario 4(b)). Overall, we note that the more uniform the dispersion of the initial state distribution across the states, the greater the computational time. A more substantial downstream effect in a larger number of states may lead to longer computational times verifying the optimality of each policy. Combining all the scenarios, a total of 20.50 million patients are excluded due to the time-limit restrictions in our sensitivity analyses.