Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

SIMULATION-BASED SETS OF SIMILAR-PERFORMING ACTIONS IN FINITE MARKOV DECISION PROCESS MODELS

Wesley J. Marrero

Thayer School of Engineering
Dartmouth College
15 Thayer Drive
Hanover, NH 03755, USA

ABSTRACT

Markov decision process (MDP) models have been used to evaluate the performance of policies in various domains, such as treatment planning in medical decision making. However, in practice, decision makers may prefer other strategies that are not statistically different from the actions in their initial policy of interest. To allow for decision makers' expertise and provide flexibility in implementing policies, this paper introduces a new framework for identifying sets of similar-performing actions in finite MDP models. These sets are obtained by combining a simulation-based dynamic programming algorithm for policy evaluation with a simulation-based statistical multiple comparisons procedure. The framework in this paper is applied in a medical decision-making setting to find sets of similar-performing antihypertensive treatment choices for a set of clinically representative patient profiles.

1 INTRODUCTION

A Markov decision process (MDP) is a mathematical model for sequential decision making under probabilistic uncertainty. MDP models have been used to inform decisions in various domains, including medicine, scheduling, transportation, finance, and energy. In many application areas, decision makers may be interested in assessing the quality of a specific policy (i.e., sequence of actions according to a decision rule). For example, a physician may be interested in evaluating the potential benefit of treating their patients according to certain clinical guidelines. This task may be accomplished with policy evaluation if the distributions of the MDP transitions are known (Puterman 2014), or temporal difference or Monte Carlo prediction if the transitions can be simulated (Sutton and Barto 2018). However, there may be situations where it is unfeasible to implement a policy or multiple policies lead to similar outcomes. Therefore, it is vital to have a framework to identify sets of actions with comparable performance to an initial policy of interest.

In settings where a human being is responsible for controlling a system, a single recommendation may not be enough, as each individual has their own decision process (Fard and Pineau 2011). It may be appropriate to assume that decision makers influence certain aspects of decision-making processes. Furthermore, the difference between the performance of the initial policy of interest and other policies may not be relevant to a decision maker. For instance, the performance of their policy of interest may not be statistically different from other policies. To test for statistical significance before observing the implications of each action in practice, I propose to simulate the effect of each action based on an estimated model of the system of interest. Decision makers can then choose between the actions in their policy of interest and similar-performing alternatives based on their expertise, preference, or other factors.

The research in decision support models that provide more than one alternative has been limited. Fard and Pineau (2011) considered the problem of generating sets of near-optimal actions for discounted infinite-horizon MDP models. While Fard and Pineau (2011) specify the difference in the performance of

978-1-6654-7661-4/22/$31.00 ©2022 IEEE

actions in the same units as the value functions, this paper defines it in terms of statistical significance. In a separate research direction, Ertefaie et al. (2016) studied the problem of providing a set of suggestions in the context of dynamic treatment regimes. Although the approach in this paper has similarities with the research of Ertefaie et al. (2016), a crucial difference is that in the present work the control is identified before the statistical inference. This difference results in fewer comparisons and improved statistical power. Another difference is that Ertefaie et al. (2016) centers on 2-stage history-dependent policies, whereas this paper focuses on Markov policies with a finite number of actions and a finite planning horizon. More recently, Tang et al. (2020) proposed a model-free algorithm that learns set-valued policies to capture near-equivalent actions. Although these studies provide sets of actions, none offer sets of similar-performing actions in the context of policy evaluation.

The ideas in this paper have been inspired by the most recent hypertension clinical guidelines from the American College of Cardiology and the American Heart Association (Whelton et al. 2018). Hypertension or high blood pressure (BP) is one of the key controllable risk factors of atherosclerotic cardiovascular disease (ASCVD), which is among the leading causes of death in the US. The latest hypertension guidelines have generated considerable controversy among practitioners, being perceived as subjective and convoluted (Ioannidis 2018; Cohen and Townsend 2018). To account for disagreement and potentially conflicting perspectives, I propose a method to enhance the clinical guidelines by allowing for physicians' expertise. This paper introduces a framework to design personalized sets of treatment choices within a margin of certainty from the actions recommended by the most recent hypertension clinical guidelines. By providing sets, the framework will account for variability in physicians' preferences, uncertainty in medications' effects, and implementation barriers.

1.1 Contributions

In previous research, my collaborators and I proposed an approach for identifying sets of near-optimal actions using a finite MDP (Marrero et al. 2021). To obtain the sets of near-optimal choices, we introduced two new algorithms. First, a simulation-based backward induction (SBBI) method that replaces the expectation in the backward induction algorithm with a sample-average approximation. Second, a simulation-based multiple comparisons with a control (SBMCC) procedure that identifies alternatives that are not statistically inferior to optimal actions. These algorithms lay the groundwork for the methods presented in this paper. Specifically, the contributions of this paper are as follows.

1. **A new simulation-based policy evaluation (SBPE) algorithm.** This algorithm replaces the approximately optimal value functions in SBBI by estimates of the value functions associated with the initial policy of interest.
2. **A new SBMCC method for policy evaluation.** This variation of the algorithm considers the case that alternatives can be better or worse than an action of interest (i.e., control).
3. **Application of the framework to find sets of similar-performing antihypertensive treatment choices.** Using a set of clinically representative patient profiles, I provide decision support that is personalized to each patient's characteristics and physicians' preferences.

Leveraging the results derived in Marrero et al. (2021), this work presents finite sample and convergence properties of the SBPE algorithm as well as convergence and asymptotic properties of SBMCC method.

1.2 Organization of the Paper

The remainder of this paper is organized as follows. Section 2 provides additional background on MDP models and multiple comparisons with a control (MCC). In Sections 3 and 4, I introduce the SBPE and SBMCC algorithms, respectively. Section 5 presents the hypertension management case study. Lastly, conclusions and future research directions are discussed in Section 6.

2 OVERVIEW AND BACKGROUND

This paper presents an approach to identify a sequence of sets of similar-performing actions to an initial policy of interest in the context of discrete-time finite MDP models. I now introduce the modeling framework, the main notions behind finite MDP models and MCC, and the mathematical notation.

Figure 1 summarizes the modeling framework for a single decision epoch of a finite MDP. I first simulate the immediate rewards and the next state transitions for each state and action at every decision epoch of the MDP. Then, I divide the outputs of the simulation model into batches and estimate action-value functions using a sample-average approximation in each batch. This method relies only on sample realizations of the transition dynamics, rather than complete knowledge of the true underlying probability distribution of the evolution of the system of interest. Subsequently, I statistically compare each action in the initial policy of interest (i.e., controls) with the remaining alternatives. The MCC method provides the strongest inference for the purposes of this paper by requiring the least number of comparisons. Once the controls are compared to each remaining alternative, the actions that perform similar to the policy of interest are identified. The sets of similar-performing actions will be referred to as sets of α-nonsignificant actions. The following subsections provide a more detailed description of MDP simulation models and MCC, as well as a formal definition of the sets of α-nonsignificant actions.

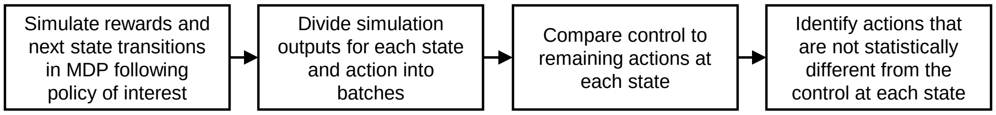
Simulate rewards and next state transitions in MDP following policy of interest
Divide simulation outputs for each state and action into batches
Compare control to remaining actions at each state
Identify actions that are not statistically different from the control at each state

Figure 1: Summary of modeling framework for a decision epoch of an MDP simulation model.

2.1 Markov Decision Process Models

An MDP model is a mathematical representation of the interactions of a decision maker with a fully observable system. The following notation will be used throughout the paper:

- t: index of discrete time periods; $t \in \mathscr{T}$, where $\mathscr{T} := \{1,\dots,T\}$ is a finite set of time periods. Decisions are made until time $T-1$; the periods $\mathscr{T} \setminus \{T\}$ will be referred to as decision epochs.
- s: state of the system; $s \in \mathscr{S}$, where $\mathscr{S} := \{1,\dots,S\}$ is a finite set of states.
- a: decision maker's action; $a \in \mathscr{A}$, where $\mathscr{A} := \{1,\dots,A\}$ is a finite set of actions.
- ω: random disturbance representing the uncertainty of the system of interest; $\omega \in \Omega$, where $\Omega := [0,1]$ is the set of all outcomes.
- $f_{t+1}(s,a,\omega)$: transition function which produces the next state s' given state s, action a, and random disturbance ω; $s' = f_{t+1}(s,a,\omega)$, where $f : \mathscr{S} \times \mathscr{A} \times \Omega \mapsto \mathscr{S}$.
- $r_t(s,a,\omega)$: random reward associated with state s, action a, and random disturbance ω, where the reward function is defined as $r : \mathscr{S} \times \mathscr{A} \times \Omega \mapsto \mathscr{R} := \{x \in \mathbb{R} | 0 \leq x < \infty\}$.
- $r_T(s,\omega)$: random terminal reward associated with state s and random disturbance ω; $r_T(s,\omega) \in \mathscr{R}$.
- γ: discount factor of the model; $\gamma \in (0,1]$.

Formally, an MDP simulation model is defined by the tuple $(\mathscr{T},\mathscr{S},\mathscr{A},f,r,\gamma)$. I assume that decision makers are interested in evaluating a policy $\pi := (\pi_t(s) : t \in \mathscr{T} \setminus \{T\}, s \in \mathscr{S})$, where $\pi_t : \mathscr{S} \mapsto \mathscr{A}$. The goal is to estimate the expected discounted reward at each decision epoch t and state s following policy π:

$$Q_t^\pi(s,\pi_t(s)) := \mathbb{E}_\omega^\pi\left[r_t(s,\pi_t(s),\omega) + \gamma Q_{t+1}^\pi\left(s',\pi_{t+1}(s')\right)|s\right],$$

where $s' = f_{t+1}(s, \pi_t(s), \omega)$ and $Q_T^\pi(s,a) := \mathbb{E}_\omega[r_T(s,\omega)|s]$ for all a. Proceeding backwards from $T-1$ until decision epoch 1, the expected discounted reward starting from each decision epoch t can be calculated.

2.2 Multiple Comparisons with a Control

When a decision maker aims to evaluate a policy, they may also be interested in comparing the performance of each action in the policy with other alternatives. In this sense, the actions in the policy serve as controls. The parameters of interest in this comparison are $Q_t^\pi(s,\pi_t(s)) - Q_t^\pi(s,a)$ for $a \in \mathscr{A}_t^-(s) := \mathscr{A} \setminus \{\pi_t(s)\}$, where $Q_t^\pi(s,a) := \mathbb{E}_\omega^\pi\left[r_t(s,a,\omega) + \gamma Q_{t+1}^\pi(\bar{s}, \pi_{t+1}(\bar{s}))|s\right]$ denotes the expected discounted reward associated with action a at decision epoch t following π afterward, and $\bar{s} = f_{t+1}(s,a,\omega)$. Given the parameters of interest, MCC intends to identify as many actions as possible to be statistically different from $\pi_t(s)$. The $1-\alpha$ simultaneous confidence intervals for the difference between a control's performance $Q_t^\pi(s,\pi_t(s))$ and the remaining action-value functions $\{Q_t^\pi(s,a) : a \in \mathscr{A}_t^-(s)\}$ are given by:

$$Q_t^\pi(s,\pi_t(s)) - Q_t^\pi(s,a) \in \hat{Q}_t^\pi(s,\pi_t(s)) - \hat{Q}_t^\pi(s,a) \pm d_t(s,\alpha)\sqrt{N^{-1}\left[\hat{\sigma}_t^2(s,\pi_t(s)) + \hat{\sigma}_t^2(s,a)\right]}, \qquad (1)$$

where $\alpha \in (0,1)$, $\hat{Q}_t^\pi(s,a)$ is an estimate of the action-value function $Q_t^\pi(s,a) < \infty$, $\hat{\sigma}_t^2(s,a)$ is an estimate of the variance of the action-value function $\sigma_t^2(s,a) < \infty$, and N is the number of observations used to calculate $\hat{Q}_t^\pi(s,a)$ and $\hat{\sigma}_t^2(s,a)$. In the case that the action-value functions are normally distributed, the quantile $d_t(s,\alpha) \in \mathbb{R}$ can be obtained using standard statistical software (Hsu 1996). But since this assumption is not satisfied in most practical situations, Westfall (2011) proposed an alternative formulation that allows for general probability distributions. Their formulation can be generalized to unequal variances by finding a quantile $d_t(s,\alpha)$ such that $\mathbb{P}\big(\max_{a\in\mathscr{A}}|\psi_t(s,a)| \leq d_t(s,\alpha)\big) = 1-\alpha$, where

$$\psi_t(s,a) := \frac{\hat{Q}_t^\pi(s,\pi_t(s)) - \hat{Q}_t^\pi(s,a) - [Q_t^\pi(s,\pi_t(s)) - Q_t^\pi(s,a)]}{\sqrt{N^{-1}\left[\hat{\sigma}_t^2(s,\pi_t(s)) + \hat{\sigma}_t^2(s,a)\right]}}, \qquad (2)$$

is a root statistic corresponding to state s and action a at decision epoch t.

2.3 Sets of Similar-Performing Actions

The implicit hypothesis in equation (1) is that all actions are equally good. Hence, if $0 \in \hat{Q}_t^\pi(s,\pi_t(s)) - \hat{Q}_t^\pi(s,a') \pm d_t(s,\alpha)\sqrt{N^{-1}\left[\hat{\sigma}_t^2(s,\pi_t(s)) + \hat{\sigma}_t^2(s,a')\right]}$, it cannot be concluded that $\pi_t(s)$ is significantly different from $a \in \mathscr{A}_t^-(s)$. This leads to our definition of a set of similar-performing actions.

Definition 1 Given N observations, a policy π, a state $s \in \mathscr{S}$, and a quantile $d_t(s,\alpha) \in \mathbb{R}$, a set of actions $\Pi_t^\pi(s,\alpha)$ is said to be *α-nonsignificant* with $\alpha \in (0,1)$ if it satisfies:

$$\Pi_t^\pi(s,\alpha) := \left\{a \in \mathscr{A} : |\hat{Q}_t^\pi(s,\pi_t(s)) - \hat{Q}_t^\pi(s,a)| \leq d_t(s,\alpha)\sqrt{N^{-1}\left[\hat{\sigma}_t^2(s,\pi_t(s)) + \hat{\sigma}_t^2(s,a)\right]}\right\}.$$

3 SIMULATION-BASED POLICY EVALUATION

This section introduces an approach to estimate action-value functions based on a policy of interest π. To estimate $Q_t^\pi(s,a)$, I introduce a policy evaluation variation of the SBBI algorithm in Marrero et al. (2021). This variation will be referred to as the SBPE algorithm, which proceeds as follows. For each decision epoch t, state s, and action a, simulate a sequence $\boldsymbol{Q}_t^\pi(s,a) := (Q_t^n(s,a,\pi) : n \in \{1,\ldots,N\})$ of $N \in \mathbb{N}\setminus\{0\}$ samples. After N observations have been simulated, approximate $Q_t^\pi(s,a)$ with its sample mean:

$$\hat{Q}_t^\pi(s,a) := \tfrac{1}{N}\textstyle\sum_{n=1}^N Q_t^n(s,a,\pi) = \tfrac{1}{N}\sum_{n=1}^N r_t(s,a,\omega^n) + \gamma\hat{Q}_{t+1}^\pi(s',\pi_{t+1}(s')),$$

where $(\omega^n : n \in \{1,\ldots,N\})$ is a sequence of independent and identically distributed (iid) random variables uniformly distributed on $[0,1]$ representing the random disturbance of the stochastic process and $s' =$

$f_{t+1}(s,a,\omega^n)$. If the terminal conditions $Q_T^{\pi}(s,a)$ are unknown, also estimate them through their sample mean. Since $f_{t+1}(s,a,\omega^n)$ and $r_t(s,a,\omega^n)$ are deterministic functions of s and a given ω^n, it follows that $\boldsymbol{Q}_t^{\pi}(s,a)$ is a sequence of iid random variables. The unknown cumulative distribution function (cdf) of $Q_t^n(s,a)$ is denoted by $\mathbb{F}_t(\cdot,s,a)$, and the joint cdf of the set $\{Q_t^n(s,a) : a \in \mathscr{A}\}$ by $\mathbb{F}_t(\cdot,s)$. Their empirical estimates are denoted by $\hat{\mathbb{F}}_t(\cdot,s,a)$ and $\hat{\mathbb{F}}_t(\cdot,s)$, respectively. Note that the SBPE algorithm is similar to Monte Carlo prediction (Sutton and Barto 2018, Chapter 5). However, the SBPE algorithm simulates the dynamics in each decision epoch independently rather than in episodes.

The SBPE algorithm has similar properties to the SBBI algorithm. For example, the algorithm converges exponentially fast on N if the rewards are bounded random variables. By Hoeffding's inequality (Hoeffding 1963, Theorem 2), it follows that:

$$\mathbb{P}\left(|\hat{Q}_t^{\pi}(s,a) - Q_t^{\pi}(s,a)| \geq \delta\right) \leq 2\exp\left\{\tfrac{-2N\delta^2}{\kappa_t^2}\right\},$$

where $\delta > 0$ and $\kappa_t := \sum_{\tau=t}^{T} \gamma^{\tau-t} R_\tau$ with $r_t(s,a,\omega) \leq R_t < \infty$ and $r_T(s,\omega) \leq R_T < \infty$. Setting the right-hand side of the inequality to less or equal than $\beta \in (0,1)$, it holds that $N \geq \kappa_t^2 \log(2/\beta)/2\delta^2$. This inequality provides the sample size required to ensure that $|\hat{Q}_t^{\pi}(s,a) - Q_t^{\pi}(s,a)| \leq \delta$ with probability of at least $1-\beta$. Furthermore, by the strong law of large numbers (Billingsley 1995, Theorem 6.1) it follows that $\hat{Q}_t^{\pi}(s,a)$ converges to $Q_t^{\pi}(s,a)$ with probability 1.

4 SIMULATION-BASED MULTIPLE COMPARISONS WITH A CONTROL

I now present the method to identify alternatives that are not statistically different from an action of interest at a significance level α. Building upon the work by Westfall (2011), the method aims to find a constant $d_t(s,\alpha)$ such that $\mathbb{P}\big(\max_{a\in\mathscr{A}} |\psi_t(s,a)| \leq d_t(s,\alpha)\big) = 1-\alpha$. The challenge in finding $d_t(s,\alpha)$ lies in its dependence on $\max_{a\in\mathscr{A}} |\psi_t(s,a)|$, which is not known because of $\{Q_t^{\pi}(s,a) : a \in \mathscr{A}\}$. The cdf of $\max_{a\in\mathscr{A}} |\psi_t(s,a)|$ is denoted by $\mathbb{H}_t(\cdot,\mathbb{F}_t(s))$, or simply $\mathbb{H}_t$ when is not necessary to highlight its dependence on $\mathbb{F}_t(\cdot,s)$. Its empirical estimate is denoted by $\hat{\mathbb{H}}_t(\cdot,\hat{\mathbb{F}}_t(s))$ or $\hat{\mathbb{H}}_t$ when there is no need to emphasize its dependence on $\hat{\mathbb{F}}_t(\cdot,s)$. The nonoverlapping batch means method is a useful tool to address the difficulty of estimating a distribution dependent on unknown parameters.

Suppose an MDP model has been simulated N times. The nonoverlapping batch means method divides the sequence of N outputs of a simulation into B adjacent nonoverlapping batches, each of size K. Since $\boldsymbol{Q}_t^{\pi}(s,a)$ is a sequence of iid random variables, dividing N outputs of an MDP simulation model into B batches is equivalent to executing B independent simulations of the MDP, each with K observations. The b^{th} batch (or simulation replicate) consists of the random variables: $Q_t^{b,1}(s,a,\pi), Q_t^{b,2}(s,a,\pi), \ldots, Q_t^{b,K}(s,a,\pi)$, for $b = 1,\ldots,B$. In each batch b, the action-value functions and their variance are estimated through the sample mean $\bar{Q}_t^b(s,a)$ and variance $\bar{\sigma}_t^2(s,a,b)$ over K observations.

Once the batching is complete, the grand sample mean $\hat{Q}_t^{\pi}(s,a)$ is attained as the average of the batch means $\bar{Q}_t^b(s,a)$ for $b = 1,\ldots,B$, and the batch sample means' variance as:

$$\hat{\zeta}_t^2(s,a) = \tfrac{1}{B-1}\sum_{b=1}^{B}\left(\bar{Q}_t^b(s,a) - \hat{Q}_t^{\pi}(s,a)\right)^2.$$

Replacing the variance of the action-value function $\sigma_t^2(s,a)$ by $K\hat{\zeta}_t^2(s,a)$ in equation (2) results in an estimator for $\psi_t(s,a)$, denoted by $\hat{\psi}_t(s,a)$.

A method to estimate the quantile $d_t(s,\alpha)$ can be developed hinging upon the adaptation of the nonoverlapping batch means method to MDP simulation models and MCC. The SBMCC algorithm for policy evaluation proceeds as follows. For each decision epoch t, state s, action a, and batch b, generate a sequence $\hat{\Psi}_t(s) := (\max_{a\in\mathscr{A}} |\bar{\psi}_t^b(s,a)| : b \in \{1,\ldots,B\})$, where $\bar{\psi}_t^b(s,a)$ is calculated as:

$$\bar{\psi}_t^b(s,a) := \frac{\bar{Q}_t^b(s,\pi_t(s)) - \bar{Q}_t^b(s,a) - \left(\hat{Q}_t^{\pi}(s,\pi_t(s)) - \hat{Q}_t^{\pi}(s,a)\right)}{\sqrt{K^{-1}\left[\bar{\sigma}_t^2(s,\pi_t(s),b) + \bar{\sigma}_t^2(s,a,b)\right]}}.$$

Using the variability across B batches, $\hat{\mathbb{H}}_t$ can be generated, and $d_t(s,\alpha)$ can be estimated as $\hat{d}_t(s,\alpha) := \inf\left\{x \in \mathbb{R} : \hat{\mathbb{H}}_t(x,\hat{\mathbb{F}}_t(s)) \geq 1-\alpha\right\} = \hat{\Psi}_t^{(i)}(s)$, where $\hat{\Psi}_t^{(1)}(s),\ldots,\hat{\Psi}_t^{(B)}(s)$ are the order statistics of $\hat{\Psi}_t(s)$, and i is chosen such that $(i-1)/B < 1-\alpha \leq i/B$.

4.1 Analysis of the SBMCC Algorithm for Policy Evaluation

This sections presents some asymptotic properties of the SBMCC algorithm for policy evaluation, assuming that $B \to \infty$ and $K \to \infty$ if $N \to \infty$. The proofs of the claims in this subsection can be found in Appendix A. Let $\Theta \subseteq \mathbb{R}$ denote the set of all possible values of $Q_t^\pi(s,\pi_t(s)) - Q_t^\pi(s,a)$. The following proposition shows that the SBMCC algorithm for policy evaluation produces the correct overall asymptotic coverage.

Proposition 1 As $N \to \infty$, it follows that:

$$\mathbb{P}\Big(Q_t^\pi(s,\pi_t(s)) - Q_t^\pi(s,a) \in \Theta : \hat{\mathbb{H}}_t\left(\max_{a\in\mathscr{A}} |\hat{\psi}_t(s,a)|, \hat{\mathbb{F}}_t(s)\right) \leq 1-\alpha\Big) = 1-\alpha.$$

This proposition suggests that the true difference between the performance of $\pi_t(s)$ and the remaining actions will asymptotically be in a subset of Θ, such that $\mathbb{P}\big(\max_{a\in\mathscr{A}} |\hat{\psi}_t(s,a)| \leq d_t(s,\alpha)\big)$ with probability of exactly $1-\alpha$. Note that although the conditions in Θ involve random variables, all the relevant quantities converge with probability 1 to their true values as $B \to \infty$ and $K \to \infty$ (see Lemma 5 in Marrero et al. (2021) and Lemma 1 in Appendix A). Building upon Proposition 1, it can be shown that a set of actions $\Pi_t^\pi(s,\alpha) \subseteq \mathscr{A}$ with a quantile $\hat{d}_t(s,\alpha)$ derived from the SBMCC algorithm will asymptotically be a set of α-nonsignificant actions with probability 1. This result is presented in the following theorem:

Theorem 1 For $\hat{d}_t(s,\alpha) = \hat{\mathbb{H}}_t^{-1}(1-\alpha,\hat{\mathbb{F}}_t(s))$, it holds that:

$$\Pi_t^\pi(s,\alpha) = \left\{a \in \mathscr{A} : |\hat{Q}_t^\pi(s,\pi_t(s)) - \hat{Q}_t^\pi(s,a)| \leq \hat{d}_t(s,\alpha)\sqrt{B^{-1}\left[\hat{\zeta}_t^2(s,\pi_t(s)) + \hat{\zeta}_t^2(s,a)\right]}\right\}$$

is a set of α-nonsignificant actions with probability 1 as $N \to \infty$.

Theorem 1 generalizes the theoretical basis of SBMCC as described in Section 5.2 of Marrero et al. (2021) to the two-sided nonparametric case.

5 CASE STUDY: PERSONALIZED HYPERTENSION TREATMENT PLANS

This section presents the application of the SBPE and SBMCC algorithms to obtain flexible hypertension treatment plans for the primary prevention of ASCVD. The sets of similar-performing actions are derived based on the 2017 Hypertension Clinical Practice Guidelines (Whelton et al. 2018), which will be referred to as the clinical guidelines throughout the rest of the paper. In the next subsections, I give some background on hypertension treatment, and describe the MDP, data source, model parameters, and modeling framework. Lastly, the treatment plans for a series of clinically representative patient profiles are presented.

5.1 Background on Hypertension Treatment

Based on the 2017 Hypertension Clinical Practice Guidelines, stage 1 hypertension is defined as a systolic blood pressure (SBP) of 130-139 mm Hg or diastolic blood pressure (DBP) of 80-89 mm Hg. An SBP of at least 140 mm Hg or a DBP of at least 90 mm Hg is considered to be stage 2 hypertension. These guidelines offer non-pharmacological and pharmacological recommendations for patients with hypertension and elevated BP, defined as an SBP of 120-129 mm Hg and a DBP smaller than 80 mm Hg. The clinical guidelines recommend pharmacological treatment for patients with stage 1 hypertension once their 10-year risk for ASCVD exceeds 10%. For patients with stage 2 hypertension, the guidelines suggest treatment until they reach controlled BP levels below stage 1 hypertension. This case study focuses on pharmacological treatment to reduce the prevalence of ASCVD before it develops (i.e., the primary prevention of ASCVD).

5.2 Markov Decision Process Formulation

This study adopts the MDP in Marrero et al. (2021). Rather than finding the treatment strategy that optimizes a certain metric (e.g., life years), the MDP is used to evaluate the effect of the clinical guidelines on patients' health. In brief, the elements of the MDP simulation model $(\mathscr{T},\mathscr{S},\mathscr{A},f,r,\gamma)$ are as follows:

- $\mathscr{T}$: 10-year planning horizon; $\mathscr{T}=\{1,\ldots,11\}$. Decisions are made at the beginning of each year $t\in\mathscr{T}\setminus\{11\}$, where $T=11$ represents the effects of treatment on patients' lifetime. This planning horizon is selected based on the major guidelines for the management of cardiovascular diseases (Whelton et al. 2018) and conversations with clinical collaborators.
- $\mathscr{S}$: state space comprising patients' demographic information (i.e., age, sex, race, smoking status), clinical observations (i.e., diabetes status, SBP, cholesterol readings), and health condition to account for their history of cardiovascular events. The state space $\mathscr{S}$ is partitioned into healthy states $\mathscr{H}$ (before adverse events) and absorbing states $\mathscr{E}$ (after adverse events), based on patients' health conditions (i.e., $\mathscr{S}=\mathscr{H}\cup\mathscr{E}$).
- $\mathscr{A}$: action space composed of 0 to 5 antihypertensive medications at a half and standard dosage, totaling $A=21$ treatment choices. This paper focuses on the number of medications since research suggests that the benefit from antihypertensive treatment is determined by the BP reduction achieved, with little effect attributable to drug-specific factors (Sundström et al. 2014).
- $f_{t+1}(s,a,\omega)$: transition function derived from patients' risk for ASCVD events (Yadlowsky et al. 2018), treatment's effect on patients' BP and risk (Sundström et al. 2014; Sussman et al. 2013), mortality from ASCVD events (NCHS 2017), and non-ASCVD mortality (Arias and Xu 2019).
- $r_t(s',a,\omega)$: reward defined as one year of perfect health minus treatment disutility; $r_t(s',a,\omega)=1-\rho(a)$ if $s'\in\mathscr{H}$ and 0 otherwise, where $\rho(a)$ denotes the disutility from medication a.
- $r_T(s',\omega)$: patients' expected lifetime after transitioning to state s'.
- γ: 3% discount of future life-year gains as recommended in Neumann et al. (2016); $\gamma=0.97$.

Please refer to Appendix B in Marrero et al. (2021) for a description of the calibration and validation of this MDP simulation model.

5.3 Patient Profiles

Based on conversations with clinical collaborators, I selected a set of patient profiles that are representative of a population with a high prevalence of ASCVD that can benefit significantly from hypertension treatment (Whelton et al. 2018). The first patient profile considered in this case study is a 40-year-old, non-diabetic, non-smoker individual with elevated BP and normal cholesterol levels. Note that this profile has no major risk factors for ASCVD events. The second and third patient profiles are identical to the first, except that their BP have reached stage 1 and stage 2 hypertension, respectively. The fourth patient profile is equal to the third, except it represents a 60-year-old patient.

To model how each patient profile's risk factors evolve over time, their progression is estimated using linear regression. Each patient profile's untreated SBP, total cholesterol, high-density lipoprotein, and low-density lipoprotein is regressed on their age, sex, race, smoking status, and diabetes status. The regression models are parameterized using data from the National Health and Nutrition Examination Survey (NHANES) from 2009 to 2014 (Centers for Disease Control and Prevention 2020). The population from NHANES is composed of adult Caucasian or African American patients from 40 to 60 years old with no recorded blood pressure treatment and no history of heart attack, stroke, or congestive heart failure.

5.4 Modeling Framework

The modeling framework for a single patient profile is summarized in Figure 2. Before developing treatment plans, the risk for ASCVD events is calculated each year. Then, the transition probabilities are estimated, and transition functions are developed. Subsequently, the treatment plan is determined based on the clinical guidelines. To derive sets of similar-performing treatment choices, I combine the SBPE and SBMCC algorithms. The treatment choices obtained from the clinical guidelines serve as controls. A significance level of $\alpha = 0.05$ is used for the statistical inference.

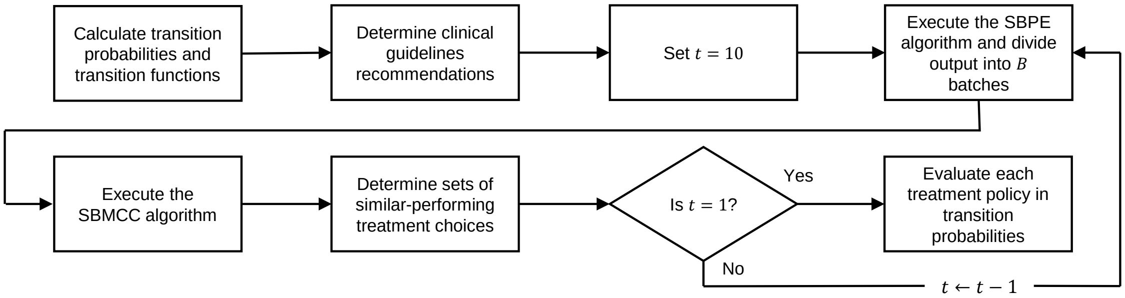
Calculate transition probabilities and transition functions
Determine clinical guidelines recommendations
Set $t = 10$
Execute the SBPE algorithm and divide output into B batches
Execute the SBMCC algorithm
Determine sets of similar-performing treatment choices
Is $t = 1$?
Yes
No
Evaluate each treatment policy in transition probabilities
$t \leftarrow t - 1$

Figure 2: Summary of modeling framework for a single patient profile.

5.5 Numerical Results

This subsection presents the recommendations of the clinical guidelines and their sets of similar-performing alternatives. As this case study concentrates on the primary prevention of ASCVD, all the results in this section correspond to patients in the healthy states $\mathscr{H}$.

5.5.1 Convergence Analysis

Before evaluating the impact of flexible treatment plans, I select the number of batches to divide the simulation output based on the maximum confidence interval half width across all patient profiles. First, the minimum number of observations K to satisfy the sample size requirements in Section 3 with $\beta = 0.01$ and $\delta = 0.5$ is fixed per batch for each patient profile. Then, the number of batches is increased iteratively until the maximum half width of the simultaneous confidence intervals across all patient profiles in the first year of the study reaches convergence. Through this method, it can be found that $B = 300$ batches may be enough to obtain a maximum confidence interval half width close to the maximum half width attained with 500 batches (Figure 3).

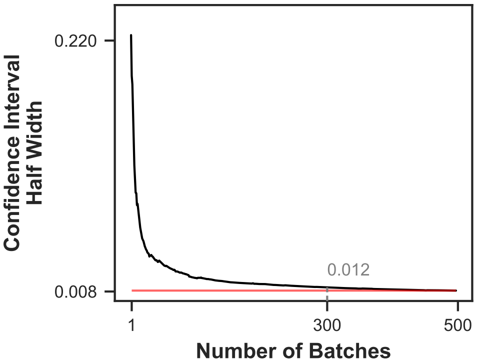

Figure 3: Convergence of the maximum confidence interval half width over the number of batches. Red line represents the maximum confidence interval half width using 500 batches (0.008).

5.5.2 Insights from Flexible Treatment

Figure 4 shows the sets of actions that perform similarly to the clinical guidelines' recommendations in the series of patient profiles. As the first patient profile has elevated BP, the clinical guidelines recommend no pharmacological treatment. However, this profile obtains sets of similar-performing actions containing 2 to 5 treatment choices. They correspond to recommending 0 to 1 medication at a standard dosage and another at half dosage over the planning horizon. These choices allow physicians and their patients to develop treatment plans that adjust to patients' tolerance to risk and medications. For example, these findings imply that the expected outcomes of a 40-year-old patient with elevated BP are not statistically different if they receive no treatment or one medication at a half dosage over the next ten years. The sets of α-nonsignificant actions identify more intense treatment choices with comparable performance to the clinical guidelines' recommendations, which may be beneficial for risk-averse patients.

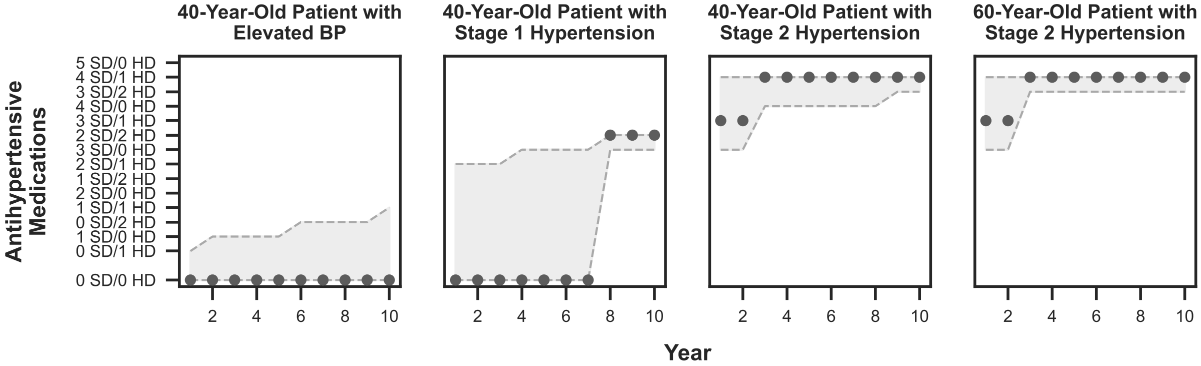

Figure 4: Sets of similar-performing treatment choices per patient profile. The sets are highlighted with the gray shaded area in each profile. The labels "SD" and "HD" denote antihypertensive medications at standard dosage and half dosage, respectively.

Increasing the profile's BP to stage 1 hypertension considerably increases the cardinality of the sets in the first seven years of the planning horizon. A reason for this is that the sets of α-nonsignificant actions consider the patient's future health, even though the clinical guidelines do not explicitly. A stage 1 hypertension BP also raises the level of aggressiveness of the treatment prescribed at year 8, when the patient exceeds a 10% 10-year risk for ASCVD events. At this point, the cardinality of the sets reduces dramatically partially due to the patient's risk increase. In the first two patient profiles, the clinical guidelines were either in the lower or upper bound of the sets in terms of the treatment recommendations.

The recommendations from the clinical guidelines increase their intensity for the third patient profile, which has a stage 2 hypertension BP level. In this profile, the sets of similar-performing actions contain substantially fewer treatment choices than in the second patient profile. Modifying the age of the third profile to 60 years old in the fourth patient profile seems to reduce the cardinality of the sets further. This behavior may be a consequence of the correlation between age and the risk for ASCVD. A higher risk may amplify the difference between the clinical guidelines' recommendations and less aggressive treatment choices. As a result, lower-intensity treatment could significantly affect the health outcomes of a patient with a high risk for ASCVD events.

6 CONCLUSIONS

This paper introduced a new framework to obtain sets of similar-performing actions in finite MDP models. It presented two algorithms to attain the sets of α-nonsignificant actions: the SBPE algorithm and the SBMCC method for policy evaluation. The SBPE algorithm was created by replacing the approximately optimal value functions in SBBI with estimates of the value functions associated with the policy being evaluated. Leveraging results in previous research, I showed that the estimates obtained with the SBPE algorithm converge to their true values with probability 1 exponentially fast on the number of observations. The SBMCC algorithm for policy evaluation was an extension of the previously developed SBMCC method to account for the possibility of alternatives that can be better or worse than the control. I proved that the algorithm reaches the correct coverage asymptotically. By offering a set of actions, the methods presented in this paper improve the usability and acceptance of MDP models in practice.

The case study studied the implications of flexible hypertension treatment plans at a patient level, based on clinically representative profiles. Two main conclusions can be made from this study. First, how much flexibility a physician may receive to treat a patient depends on the patient's characteristics (e.g., age and BP levels). In general, patients with higher risk for ASCVD events obtain fewer treatment choices in the sets than patients with lower risk. Second, the sets contain monotonic treatment choices in the number of medications. The reason for this is that the natural ordering of state and action spaces in the case study satisfies the conditions established in Theorem 3 of Marrero et al. (2021). If these conditions are not satisfied, the actions in the sets may not follow an intuitive order to decision makers. As evidence of the effectiveness of BP treatment becomes increasingly available, the sets of α-nonsignificant antihypertensive medications may become more accurate in medical practice.

There are opportunities for future work that build upon the sets of similar-performing actions. From a technical perspective, this research could be extended to policy evaluation in partially observable and infinite-horizon MDP models. Ideas from Haskell et al. (2016) could be used to obtain empirical estimates of the value and action-value functions for the latter type of models. The SBPE and SBMCC algorithms for policy evaluation inherit some of the curses of dimensionality associated with standard policy evaluation. Overcoming these challenges could be another area for future work. Also, the algorithms are limited by their storage requirements. This limitation could be addressed by designing an online method to obtain sets of α-nonsignificant actions. From a clinical point of view, this work can be extended by incorporating other conditions, like high cholesterol or diabetes. In addition, the methods in this paper could be expanded to allow for multiple scenarios of the input data, which may be helpful when there is disagreement on how to model patients' health progression.

The algorithms in this paper continue a new line of work that handles stochastic optimization problems as hypothesis testing problems to provide decision makers with flexibility. Having multiple alternatives in sequential decision problems offer domain experts an effective way to integrate their knowledge into mathematical models. The sets of similar-performing choices could have many benefits in practice, including flexibility and better user experience.

ACKNOWLEDGMENTS

Special thanks to Mariel S. Lavieri and Ambuj Tewari for their input in the derivation of the SBBI and one-sided SBMCC algorithms, which established the foundation for the methods introduced in this paper. I am also grateful to Jeremy B. Sussman and Rodney A. Hayward for their guidance and support in the medical components of this work, including the development of the clinically representative patient profiles. Lastly, I would like to thank Gian-Gabriel P. Garcia for his help brainstorming the initial ideas of the two-sided SBMCC algorithm.

A PROOF OF ANALYTICAL RESULTS

The proof of Proposition 1 depends on the following lemma:

Lemma 1 $\sup_{x\in\mathbb{R}} |\hat{\mathbb{H}}_t(x,\hat{\mathbb{F}}_t(s)) - \mathbb{H}_t(x,\mathbb{F}_t(s))|$ converges to 0 almost surely (a.s.).

Proof. As $\boldsymbol{Q}_t^{\pi}(s,a)$ is a sequence of iid random variables, it follows that $\sup_{x\in\mathbb{R}} |\hat{\mathbb{F}}_t(x,s,a) - \mathbb{F}_t(x,s,a)| \overset{a.s.}{\to} 0$ by the Glivenko-Cantelli Theorem (Billingsley 1995, Theorem 20.6). Since $Q_t^n(s,a,\pi)$ is independent from

$Q_t^n(s,a',\pi)$ for $a \neq a'$, $\mathbb{F}_t(\cdot,s) = \prod_{a\in\mathscr{A}} \mathbb{F}_t(\cdot,s,a)$ and $\hat{\mathbb{F}}_t(\cdot,s) = \prod_{a\in\mathscr{A}} \hat{\mathbb{F}}_t(\cdot,s,a)$. Because $\hat{\mathbb{F}}_t(\cdot,s,a) \overset{a.s.}{\to} \mathbb{F}_t(\cdot,s,a)$ uniformly on $\mathbb{R}$, it also holds that $\hat{\mathbb{F}}_t(\cdot,s) \overset{a.s.}{\to} \mathbb{F}_t(\cdot,s)$ uniformly on $\mathbb{R}^A$. As $\hat{Q}_t^\pi(s,a) \overset{a.s.}{\to} Q_t^\pi(s,a)$, $\hat{\sigma}_t^2(s,a) \overset{a.s.}{\to} \sigma_t^2(s,a)$, and $K\hat{\zeta}_t^2(s,a) \overset{a.s.}{\to} \sigma_t^2(s,a)$ by the strong law of large numbers and Lemmas 2 and 5 in Marrero et al. (2021), $\lim\limits_{B\to\infty}\lim\limits_{K\to\infty} \hat{\mathbb{H}}_t(\cdot,\hat{\mathbb{F}}_t(s)) = \lim\limits_{N\to\infty} \mathbb{H}_t(\cdot,\mathbb{F}_t(s))$.

Now, notice that $\mathbb{H}_t(x,\mathbb{F}_t(s)) = \mathbb{P}\left(\max_{a\in\mathscr{A}} |\psi_t(s,a)| \le x\right) = \mathbb{P}\left(-x \le \psi_t(s,a^1) \le x, \ldots, -x \le \psi_t(s,a^A) \le x\right)$ for $a^1,\ldots,a^A \in \mathscr{A}$. From the Central Limit Theorem (Casella and Beroer 2001, Theorem 5.5.14), it holds that $\lim\limits_{N\to\infty} \mathbb{P}\left(-x \le \psi_t(s,a) \le x\right) = \Phi(x) - \Phi(-x) = 2\Phi(x) - 1$ for $x \ge 0$ and 0 for $x < 0$, where $\Phi(\cdot)$ is the cdf of a standard normal random variable. Thus, the individual components of the joint cdf are continuous in the limit. As the composite of continuous functions is continuous (Bartle and Sherbert 2010, Theorem 5.2.7), $\lim\limits_{N\to\infty} \mathbb{H}_t(x,\mathbb{F}_t(s))$ is continuous. Further, by Pólya's Theorem (Serfling 1980, Theorem 1.5.3) it follows that $\lim\limits_{N\to\infty} \sup_{x\in\mathbb{R}} \left|\hat{\mathbb{H}}_t(x,\hat{\mathbb{F}}_t(s)) - \mathbb{H}_t(x,\mathbb{F}_t(s))\right| = 0$ due to the continuity of $\lim\limits_{N\to\infty} \mathbb{H}_t(x,\mathbb{F}_t(s))$ for any $x \in \mathbb{R}$. Merging this result with the conclusion that $\hat{\mathbb{F}}_t(\cdot,s) \overset{a.s.}{\to} \mathbb{F}_t(\cdot,s)$ uniformly, it follows that $\sup_{x\in\mathbb{R}} \left|\hat{\mathbb{H}}_t(x,\hat{\mathbb{F}}_t(s)) - \mathbb{H}_t(x,\mathbb{F}_t(s))\right| \overset{a.s.}{\to} 0$. □

Proof of Proposition 1. From Lemma 1, it follows that $\sup_{x\in\mathbb{R}} |\hat{\mathbb{H}}_t(x,\hat{\mathbb{F}}_t(s)) - \mathbb{H}_t(x,\mathbb{F}_t(s))| \overset{a.s.}{\to} 0$. Because a.s. convergence implies convergence in distribution, it holds that $\sup_{x\in\mathbb{R}} |\hat{\mathbb{H}}_t(x,\hat{\mathbb{F}}_t(s)) - \mathbb{H}_t(x,\mathbb{F}_t(s))| \overset{\mathscr{D}}{\to} 0$. Since $\mathbb{H}_t(\cdot,\mathbb{F}_t(s))$ is a true cdf, it is uniformly distributed on $[0,1]$, denoted by $\mathscr{U}(0,1)$. Consequently, $\hat{\mathbb{H}}_t(\cdot,\hat{\mathbb{F}}_t(s))$ must follow a $\mathscr{U}(0,1)$ distribution asymptotically. Hence, $\mathbb{P}(\hat{\mathbb{H}}_t(\cdot,\hat{\mathbb{F}}_t(s)) \le 1-\alpha) = \mathbb{P}(\mathscr{U}(0,1) \le 1-\alpha) = 1-\alpha$ as $N \to \infty$. □

Proof of Theorem 1. By Proposition 1, it follows that:

$$\begin{aligned}
&\mathbb{P}\Big(Q_t^\pi(s,\pi_t(s)) - Q_t^\pi(s,a) \in \Theta : \hat{\mathbb{H}}_t\left(\max_{a\in\mathscr{A}} |\hat{\psi}_t(s,a)|, \hat{\mathbb{F}}_t(s)\right) \le 1-\alpha\Big) = \mathbb{P}\Big(\max_{a\in\mathscr{A}} |\hat{\psi}_t(s,a)| \le \hat{d}_t(s,\alpha)\Big) \\
&= \mathbb{P}\Big(\hat{Q}_t^\pi(s,\pi_t(s)) - \hat{Q}_t^\pi(s,a^1) - \Delta \le Q_t^\pi(s,\pi_t(s)) - Q_t^\pi(s,a^1) \le \hat{Q}_t^\pi(s,\pi_t(s)) - \hat{Q}_t^\pi(s,a^1) + \Delta, \ldots, \\
&\quad \hat{Q}_t^\pi(s,\pi_t(s)) - \hat{Q}_t^\pi(s,a^A) - \Delta \le Q_t^\pi(s,\pi_t(s)) - Q_t^\pi(s,a^A) \le \hat{Q}_t^\pi(s,\pi_t(s)) - \hat{Q}_t^\pi(s,a^A) + \Delta\Big) = 1-\alpha,
\end{aligned}$$

as $N \to \infty$, where $\Delta := \hat{d}_t(s,\alpha)\sqrt{B^{-1}\left[\hat{\zeta}_t(s,\pi_t(s)) + \hat{\zeta}_t^2(s,a)\right]}$ and $a^1,\ldots,a^A \in \mathscr{A}$. Furthermore, from Lemma 1 it holds that $\hat{\mathbb{H}}_t(1-\alpha,\hat{\mathbb{F}}_t(s)) \overset{a.s.}{\to} \mathbb{H}_t(1-\alpha,\mathbb{F}_t(s))$. Therefore, $\hat{d}_t(s,\alpha) \overset{a.s.}{\to} d_t(s,\alpha)$ by Theorem 2.3.1 in Serfling (1980). Consequently, the asymptotic confidence intervals simultaneously contain $Q_t^\pi(s,\pi_t(s)) - Q_t^\pi(s,a^1), \ldots, Q_t^\pi(s,\pi_t(s)) - Q_t^\pi(s,a^A)$ with probability exactly $1-\alpha$. Moreover, under the implicit null hypothesis that all actions have the same performance, any action a such that $|\hat{Q}_t^\pi(s,\pi_t(s)) - \hat{Q}_t^\pi(s,a)| \le \Delta$ is not statistically significant from $\pi_t(s)$. □

REFERENCES

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

Bartle, R. G., and D. R. Sherbert. 2010. *Introduction to Real Analysis*. 4 ed. Hoboken, NJ, USA: John Wiley & Sons.

Billingsley, P. 1995. *Probability and Measure Theory*. New York, NY, USA: John Wiley and Sons.

Casella, G., and R. Beroer. 2001. *Statistical lnference*. 2 ed. Belmont, California: Duxbury Press.

Centers for Disease Control and Prevention 2020. "National Health and Nutrition Examination Survey Data".

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

Ertefaie, A., T. Wu, K. G. Lynch, and I. Nahum-Shani. 2016. "Identifying a Set That Contains the Best Dynamic Treatment Regimes". *Biostatistics* 17(1):135–148.

Fard, M. M., and J. Pineau. 2011. "Non-deterministic Policies in Markovian Decision Processes". *Journal of Artificial Intelligence Research* 40:1–24.

Haskell, W. B., R. Jain, and D. Kalathil. 2016. “Empirical Dynamic Programming”. *Mathematics of Operations Research* 41(2):402–429.

Hoeffding, W. 1963. “Probability Inequalities for Sums of Bounded Random Variables”. *Journal of the American Statistical Association* 58(301):13–30.

Hsu, J. 1996. *Multiple Comparisons: Theory and Methods*. London, UK: CRC Press.

Ioannidis, J. P. 2018. “Diagnosis and Treatment of Hypertension in the 2017 Acc/Aha Guidelines and in the Real World”. *JAMA - Journal of the American Medical Association* 319(2):115–116.

Marrero, W. J., M. S. Lavieri, J. B. Sussman, and R. A. Hayward. 2021. “Data-Driven Ranges of Near-Optimal Actions for Finite Markov Decision Processes”. *Optimization Online*:1–60.

NCHS 2017. “Health, United States, 2016: with Chartbook on Long-Term Trends in Health”. *Center for Disease Control*:314–317.

Neumann, P., G. Sanders, L. Russell, and J. Siegel. 2016. *Cost-Effectiveness in Health and Medicine*. New York, NY, USA: Oxford University Press.

Puterman, M. L. 2014. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ, USA: John Wiley & Sons.

Serfling, R. J. 1980. *Approximation Theorems of Mathematical Statistics*, Volume 145 of *Wiley Series in Probability and Statistics*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Sundström, J., H. Arima, M. Woodward, R. Jackson, K. Karmali, D. Lloyd-Jones, C. Baigent, J. Emberson, K. Rahimi, S. Macmahon, A. Patel, V. Perkovic, F. Turnbull, B. Neal, L. Agodoa, R. Estacio, R. Schrier, J. Lubsen, J. Chalmers, J. Cutler, B. Davis, L. Wing, N. R. Poulter, P. Sever, G. Remuzzi, P. Ruggenenti, S. Nissen, L. H. Lindholm, T. Fukui, T. Ogihara, T. Saruta, H. Black, P. Sleight, M. Lievre, H. Suzuki, K. Fox, L. Lisheng, T. Ohkubo, Y. Imai, S. Yusuf, C. J. Bulpitt, E. Lewis, M. Brown, C. Palmer, J. Wang, C. Pepine, M. Ishii, Y. Yui, K. Kuramoto, M. Pfeffer, F. W. Asselbergs, W. H. van Gilst, B. Byington, B. Pitt, B. Brenner, W. J. Remme, D. de Zeeuw, M. Rahman, G. Viberti, K. Teo, A. Zanchetti, E. Malacco, G. Mancia, J. Staessen, R. Fagard, R. Holman, L. Hansson, J. Kostis, Y. Kanno, S. Lueders, M. Matsuzaki, P. Poole-Wilson, J. Schrader, K. Rahimi, C. Anderson, J. Chalmers, N. Chapman, R. Collins, B. Neal, A. Rodgers, P. Whelton, M. Woodward, and S. Yusuf. 2014. “Blood PressurE-lowering Treatment Based on Cardiovascular Risk: A Meta-Analysis of Individual Patient Data”. *The Lancet* 384(9943):591–598.

Sussman, J., S. Vijan, and R. Hayward. 2013. “Using Benefit-Based Tailored Treatment to Improve the Use of Antihypertensive Medications”. *Circulation* 128(21):2309–2317.

Sutton, R. S., and A. G. Barto. 2018. *Reinforcement Learning: An Introduction*. 2 ed. Cambridge, MA, USA: MIT press.

Tang, S., A. Modi, M. W. Sjoding, and J. Wiens. 2020. “Clinician-In-ThE-loop Decision Making: Reinforcement Learning with Near-Optimal Set-Valued Policies”. 9329–9338. Virtual Event: 37th International Conference on Machine Learning.

Westfall, P. H. 2011. “On Using the Bootstrap for Multiple Comparisons”. *Journal of Biopharmaceutical Statistics* 21(6):1187–1205.

Whelton, P. K., R. M. Carey, W. S. Aronow, D. E. Casey, K. J. Collins, C. Dennison Himmelfarb, S. M. DePalma, S. Gidding, K. A. Jamerson, D. W. Jones, E. J. MacLaughlin, P. Muntner, B. Ovbiagele, S. C. Smith, C. C. Spencer, R. S. Stafford, S. J. Taler, R. J. Thomas, K. A. Williams, J. D. Williamson, and J. T. Wright. 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.

Yadlowsky, S., R. A. Hayward, J. B. Sussman, R. L. McClelland, Y. I. Min, and S. Basu. 2018. “Clinical Implications of Revised Pooled Cohort Equations for Estimating Atherosclerotic Cardiovascular Disease Risk”. *Annals of Internal Medicine* 169(1):20–29.

AUTHOR BIOGRAPHIES

WESLEY J. MARRERO is an Assistant Professor of Engineering in Thayer School at Dartmouth College. Before joining Dartmouth, he was a postdoctoral research fellow in the Massachusetts General Hospital Institute for Technology Assessment at Harvard Medical School. He earned his Ph.D. in the Department of Industrial and Operations Engineering at the University of Michigan. His research interests lie at the intersection of operations research and statistics with an emphasis on stochastic simulation and optimization to support decision making in practice. His current work addresses various application areas, including substance use disorder, cardiovascular disease, and organ transplantation. His e-mail address is wesley.marrero@dartmouth.edu. His website is https://engineering.dartmouth.edu/community/faculty/wesley-marrero.