Robust Markov Decision Processes for Medical Treatment Decisions

Yuanhui Zhang
Operations Research Program, North Carolina State University, yuanhui.zhang@gmail.com

Lauren N. Steimle
Department of Industrial and Operations Engineering, University of Michigan, steimle@umich.edu

Brian T. Denton
Department of Industrial and Operations Engineering, University of Michigan, btdenton@umich.edu

Medical treatment decisions involve complex tradeoffs between the risks and benefits of various treatment options. The diversity of treatment options that patients can choose over time and uncertainties in future health outcomes, result in a difficult sequential decision making problem. Markov decision processes (MDPs) are commonly used to study medical treatment decisions; however, optimal policies obtained by solving MDPs may be affected by the uncertain nature of the model parameter estimates. In this article, we present a robust Markov decision process treatment model (RMDP-TM) with an uncertainty set that incorporates an uncertainty budget into the formulation for the transition probability matrices (TPMs) of the underlying Markov chain. We show that the addition of an uncertainty budget can control the tradeoff between mean performance and worst-case performance of the resulting policies. Further, we present theoretical analysis to establish computationally efficient methods to solve the RMDP-TM and we provide conditions under which the policy of nature is stationary. Finally, we present an application of the models to a medical treatment decision problem of optimizing the sequence and the start time to initiate medications for glycemic control for patients with type 2 diabetes.

Key words: robust optimization; robust Markov decision process; medical decision making; type 2 diabetes; glycemic control

Subject classifications: Dynamic programming/optimal control: Application. Dynamic programming/optimal control: Markov: Finite states. Health care: Treatment.

Area of review: Applications in Biology, Medicine, & Health Care

1. Introduction

Medical treatment decisions for chronic diseases involve complex tradeoffs between benefits and harms of treatment. These decisions are often made by physicians based on results from randomized control trials and/or observational studies. However, decisions are rarely quantified in a way that makes clear the short term costs of medications and harms from medication side effects versus the long term benefits of avoiding disease-related complications. For many chronic diseases, there are multiple treatment options that can potentially be selected over the course of a patient's lifetime. This results in a difficult sequential decision making problem the goal of which is to optimally trade off short term harms (e.g. side effects, medication costs) with uncertain long term benefits (e.g. delaying or avoiding disease-related complications or death).

Markov decision process (MDP) models have been used to study optimal control of many medical treatment decisions including liver transplants, HIV, diabetes, and others (Alagoz et al. 2004, 2007, Shechter et al. 2008, Mason et al. 2014). A key component of every MDP model is the transition probability matrix (TPM) which describes stochastic changes in the system over time. The optimal policy of an MDP may be highly dependent on TPMs of the underlying Markov chain (Mannor et al. 2016). A strong assumption of MDP models is that TPMs are known with certainty. For medical treatment decisions, maximum likelihood estimates (MLEs) of the TPMs from a population-based observational data are commonly used; however, MLEs cannot capture the natural variation in transition probabilities caused by patient heterogeneity. Therefore, it is potentially valuable to develop optimization models which can also take into account this variation.

Robust MDPs (RMDPs), presented in a max-min framework, have been developed to generate the worst-case optimal policy when MDP parameters such as TPMs are subject to uncertainty (Satia and Lave Jr. 1973, White III and Eldeib 1994, Kaufman and Schaefer 2012, Iyengar 2005, Nilim and El Ghaoui 2005, Wiesemann et al. 2013). However, the application of RMDPs to medical decision making problems is very limited. In this article, we propose a finite-horizon RMDP treatment model (RMDP-TM) with a controllable uncertainty set formulation for the uncertain TPMs for optimizing medical treatment decisions. Similar to the RMDP framework presented in Nilim and El Ghaoui (2005), and Iyengar (2005), the uncertainty sets for TPMs are assumed to satisfy the *rectangular uncertainty property*. In order to control the conservativeness of the robust optimal policy, we include

an *uncertainty budget* as part of the uncertainty set formulation to control the size of the uncertainty set.

We present theoretical analysis to establish computationally efficient methods to solve the RMDP-TM, and to provide sufficient conditions under which the RMDP-TM is equivalent to its (often more difficult) time-invariant counterpart in which the TPMs are constrained to be the same in all decision epochs. We use the RMDP-TM version of the glycemic control model, based on our previously published paper (Zhang et al. 2014), to demonstrate the application of the proposed model to the context of optimizing treatment decisions for patients with type 2 diabetes. We use this example to demonstrate structural differences between the RMDP-TM optimal policy and the MDP optimal policy, to compare the nominal and worst-case performance of various robust optimal treatment policies, as confidence level and uncertainty budget changes, and to draw conclusions about the feasibility and the value of using the proposed RMDP-TM for real-world medical treatment decision making problems.

The main contributions of this article are twofold: from the methodological point of view: (1) we present a new robust stochastic optimization model, the RMDP-TM, which is suited to medical treatment decision problems that arise in the context of chronic diseases; (2) we propose an easy-to-implement uncertainty set formulation for TPMs, which can be used to control the conservativeness of the optimal solution; (3) we present computationally efficient methods for solving this model; and (4) we prove that the time-invariant version of the RMDP-TM can be solved in polynomial time under certain conditions.

From the application point of view, (1) we present a case study applying the RMDP framework to a real-world medical treatment decision problem in the context of glycemic control for patients with type 2 diabetes; (2) we utilize a validated glycemic control model with parameters estimated from a large longitudinal data set comprised of millions of patients' medical records and pharmacy claims; (3) we illustrate the structure and performance of various robust optimal treatment policies and the MDP optimal policy which can potentially be valuable to clinicians and policy makers in better understanding model-informed treatment decisions.

The remainder of this paper is organized as follows. In Section 2 we present the mathematical formulation of the RMDP-TM for general medical treatment decision problems. In Section 3, we present theoretical analysis and solution methods for solving the RMDP-TM.

In Section 4 we demonstrate the application of the RMDP-TM to optimize the treatment decisions with regard to the sequence and time to initiate hyperglycemia lowering medications for patients with type 2 diabetes. Finally, in Section 5 we highlight our main conclusions.

2. Model Formulation

We formulate the medical treatment decision process as a finite-state, finite-horizon, discrete-time MDP where the underlying Markov chain represents the progression of a patient's health status. The model includes the following five components:

Time horizon. We assume treatment decisions are made at a finite and discrete set of time epochs indexed by $t \in \mathcal{T} = \{1, 2, \ldots, T\}$ where 1 represents the time of being diagnosed with a certain disease, and T represents a reasonable upper limit on patients' age (e.g. age 100). The period after the last epoch T is called the *post decision horizon*.

States. The model includes a discrete set of *health states*, $\mathcal{L}$, a discrete set of *treatment states*, $\mathcal{M}$, and an *absorbing state*, $\mathcal{D}$. The health state, $\ell_t \in \mathcal{L} = \{\ell(1), \ldots, \ell(|\mathcal{L}|)\}, \forall t \in \mathcal{T}$, represents the severity of the disease at time epoch t. The treatment state, $\boldsymbol{m}_t = (m_{1,t}, m_{2,t}, \ldots, m_{n,t}) \in \mathcal{M}$, is an $n-$tuple binary vector in which n represents the total number of available treatment options (e.g. available medications); $m_{i,t} = 1, \forall i \in \{1, \ldots, n\}$, represents the patient is on treatment option i at the beginning of the time epoch t, otherwise, $m_{i,t} = 0$. Health states and treatment states influence the risk of disease-related complications that the treatment aims to prevent. The absorbing state, $\mathcal{D}$, is commonly included in medical treatment decision models (Alagoz et al. 2004, 2007, Shechter et al. 2008, Kurt et al. 2011, Mason et al. 2014). It includes all major disease-related complications that the decision maker aims to prevent (e.g. heart attack, stroke, and renal failure) and death from any causes. The complete set of states in the model is given by $\mathcal{S} = \{\mathcal{L} \times \mathcal{M}\} \bigcup \{\mathcal{D}\}$. Note that although $\mathcal{L}$ and $\mathcal{M}$ are defined independently, they are inter-dependent due to the effect of treatment actions on the probability of entering the absorbing state.

Actions. The action at time epoch $t \in \mathcal{T} \backslash \{T\}$ is denoted by $\boldsymbol{\alpha}_t(\ell_t, \boldsymbol{m}_t) \in \mathcal{A}_t$, and it represents the selection of treatment option(s) to initiate during the time period $(t, t+1]$ given the patient is in health state $\ell_t \in \mathcal{L}$, and treatment state $\boldsymbol{m}_t \in \mathcal{M}$. We adopt the convention that, in finite horizon problems, decisions are not made at the last time epoch T. The set of all possible actions at time epoch t is denoted by $\mathcal{A}_t = \{(A_{1,t}, A_{2,t}, \ldots, A_{n,t}) |\, A_{i,t} \in$

$\{I, D\}, \forall i \in \{1, \ldots, n\}\}$ where $A_{i,t} = I$, $\forall i \in \{1, \ldots, n\}$, represents to initiate treatment option i at time epoch t, otherwise $A_{i,t} = D$. For a patient in the absorbing state, there are no further actions, i.e., $\mathcal{A}_t = \emptyset$, $t \in \mathcal{T} \backslash \{T\}$. We assume that no future treatment decision will be made during the post decision horizon, therefore the treatment state is assumed to be the same as $\boldsymbol{m}_T$ during the post decision horizon. For ease of expression, we simplify the notation $\boldsymbol{\alpha}_t(\ell_t, \boldsymbol{m}_t)$ to $\boldsymbol{\alpha}_t(s_t)$. Given a patient in treatment state, $\boldsymbol{m}_t$, at the beginning of epoch t, and take the action, $\boldsymbol{\alpha}_t(s_t)$, we denote $\boldsymbol{m}_{t+1}(\boldsymbol{\alpha}_t(s_t))$ to be the treatment state at the beginning of the next time epoch $t+1$. A *decision rule*, $d_t : \mathcal{S} \to \mathcal{A}_t, \forall t \in \mathcal{T} \backslash \{T\}$, specifies the action when the patient is in state $s_t \in \mathcal{S}$ at time epoch $t \in \mathcal{T} \backslash \{T\}$.

Reward. The *immediate reward* is denoted by $r_t(s_t, \boldsymbol{\alpha}_t(s_t)), \forall t \in \mathcal{T} \backslash \{T\}$, and it represents the reward received during the time period $(t, t+1]$ given being in state s_t and taking action $\boldsymbol{\alpha}_t(s_t)$ at time epoch t. The *terminal reward*, which defines the boundary condition of the model, is denoted by $r_T(s_T), \forall s_T \in \mathcal{S}$, and it represents the expected total reward accumulated during the post decision horizon.

Probabilities. There are two types of transition probabilities in the model: probabilities of entering the absorbing state and probabilities of transitioning among health states. We denote the probability of entering the absorbing state as $p_t^{\mathrm{E}}(s_t, \boldsymbol{\alpha}_t(s_t))$. For any $s_t \in \mathcal{L} \times \mathcal{M}$, $p_t^{\mathrm{E}}(s_t, \boldsymbol{\alpha}_t(s_t))$ represents the probability of having at least one disease-related complication or death occurs during the time period $(t, t+1]$. For $s_t = \mathcal{D}$, $p_t^{\mathrm{E}}(s_t, \boldsymbol{\alpha}_t(s_t))$ represents the probability of staying in the absorbing state which equals 1. We denote the transition probability between health states conditional on not entering the absorbing state during the time period $(t, t+1]$ by $q_{t,\ell_t}(\ell_{t+1}), \forall t \in \mathcal{T} \backslash \{T\}, \ell_t, \ell_{t+1} \in \mathcal{L}$. The probability of transitioning from state $s_t \in \mathcal{S}$ to state $s_{t+1} \in \mathcal{S}$, given that the action, $\boldsymbol{\alpha}_t(s_t) \in \mathcal{A}_t$, is taken at time epoch t, is defined as follows:

$$p_t(s_{t+1}|s_t, \boldsymbol{\alpha}_t(s_t)) = \begin{cases} (1 - p_t^{\mathrm{E}}(s_t, \boldsymbol{\alpha}_t(s_t)))q_{t,\ell_t}(\ell_{t+1}), & \text{if } s_t, s_{t+1} \in \mathcal{L} \times \mathcal{M}, \\ p_t^{\mathrm{E}}(s_t, \boldsymbol{\alpha}_t(s_t)), & \text{if } s_t \in \mathcal{L} \times \mathcal{M}, s_{t+1} = \mathcal{D}, \\ 1, & \text{if } s_t = \mathcal{D}, s_{t+1} = \mathcal{D} \\ 0, & \text{if } s_t = \mathcal{D}, s_{t+1} \in \mathcal{L} \times \mathcal{M}. \end{cases}$$

2.1. MDP formulation

A *policy*, $\boldsymbol{\pi} = \{d_1, d_2, \ldots, d_{T-1}\}$, is a sequence of decision rules that specifies the action to be used at each state in each time epoch. The policy induces a probability distribution

on the set of all realizations of the MDP. Based on the expected total discounted reward criterion, the value function for a given policy $\boldsymbol{\pi}$ is defined as follows: ==

$$v_t^{\boldsymbol{\pi}}(s_t) \equiv \mathbb{E}_{s_t}^{\boldsymbol{\pi}}\left[\sum_{k=t}^{T-1} \lambda^{k-t} r_k(s_k, \boldsymbol{\alpha}_k(s_k)) + \lambda^{T-t} r_T(s_T)\right], \forall s_t \in \mathcal{S},\, t \in \mathcal{T}. \tag{1}$$

where $v_t^{\boldsymbol{\pi}}(s_t)$ represents the expected total discounted reward accumulated from time epoch t onward. Based on the definition of the value function for a given policy, the optimal value function for the MDP can be written as follows:

$$v_t^{\mathrm{MDP}}(s_t) = \begin{cases} \max\limits_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} \mathbb{E}_{s_t}^{\boldsymbol{\pi}}\left[\sum\limits_{k=t}^{T-1} \lambda^{k-t} r_k(s_k, \boldsymbol{\alpha}_k(s_k)) + \lambda^{T-t} r_T(s_T)\right], \forall\, s_t \in \mathcal{L}\times\mathcal{M}, \\ 0, \qquad\qquad \text{otherwise}, \end{cases} \quad \forall t \in \mathcal{T}, \tag{2}$$

where $\boldsymbol{\Pi}$ represents the set of all possible policies. It is well-known that the optimal value function of the MDP (2) can be rewritten recursively as follows:

$$v_t^{\mathrm{MDP}}(s_t) = \begin{cases} \max\limits_{\boldsymbol{\alpha}_t(s_t)\in\mathcal{A}_t} \Big\{ r_t(s_t, \boldsymbol{\alpha}_t(s_t)) + \lambda(1 - p_t^{\mathrm{E}}(s_t, \boldsymbol{\alpha}_t(s_t))) \sum\limits_{\ell_{t+1}\in\mathcal{L}} q_{t,\ell_t}(\ell_{t+1}) \\ \times v_{t+1}^{MDP}(\ell_{t+1}, \boldsymbol{m}_{t+1}(\boldsymbol{\alpha}_t(s_t))) \Big\}, \forall s_t = (\ell_t, \boldsymbol{m}_t) \in \mathcal{L}\times\mathcal{M}, t \in \mathcal{T}\backslash\{T\}, \\ r_T(s_T), \qquad\qquad \forall s_T = (\ell_T, \boldsymbol{m}_T) \in \mathcal{L}\times\mathcal{M}, \\ 0, \qquad\qquad \forall s_t \in \mathcal{D}, \end{cases} \quad \forall t \in \mathcal{T}, \tag{3}$$

where $v_t^{\mathrm{MDP}}(s_t)$ denotes the optimal value-to-go. The backward induction algorithm presented in Puterman (1994) is commonly used to solve MDP models. The MDP formulation assumes no uncertainty in the health state TPM, $Q_t \triangleq [q_{\ell_t}(\ell_{t+1})]_{|\mathcal{L}|\times|\mathcal{L}|}$, so we refer to it as the *nominal TPM*, and denoted it by $\hat{Q}_t \triangleq [\hat{q}_{\ell_t}(\ell_{t+1})]_{|\mathcal{L}|\times|\mathcal{L}|}$, from hereafter.

2.2. RMDP-TM formulation

In many medical treatment decision problems, the health state TPM, Q_t, is estimated from a longitudinal data set by using the maximum likelihood method (Craig and Sendi 2002); therefore, it is subject to uncertainty caused by sample variation. We assume that the health state TPM lies in a given uncertainty set $\mathcal{Q}_t$ which has the rectangular uncertainty property, i.e. $\mathcal{Q}_t = \prod\limits_{\ell_t\in\mathcal{L}} \mathcal{Q}_{t,\ell_t}$ where $\mathcal{Q}_{t,\ell_t}$ denotes the uncertainty set of row ℓ_t of the TPM, Q_t. The rectangular uncertainty property indicates the choice of the transition probabilities when the system is in state $\ell_t \in \mathcal{L}$ at epoch t, is independent of the choice of the transition probabilities when the system is in state $\ell_t' \neq \ell_t \in \mathcal{L}$ at epoch t. This property is key to maintaining the tractability of solving the RMDPs.

We assume the goal of the RMDP-TM is to maximize the worst-case expected total discounted reward. The optimal value function of the RMDP-TM can be written as follows:

$$v_t^{\text{RMDP-TM}}(s_t) = \begin{cases} \max\limits_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} \min\limits_{\boldsymbol{\theta}_t\in\Theta_t} \mathbb{E}_{s_t}^{\boldsymbol{\pi},\boldsymbol{\theta}_t}\left[\sum\limits_{k=t}^{T-1} \lambda^{k-t} r_k(s_k, \boldsymbol{\alpha}_k(s_k)) + \lambda^{T-t} r_T(s_T)\right], & \forall\, s_t \in \mathcal{L}\times\mathcal{M}, \\ 0, & \forall s_t \in \mathcal{D}, \end{cases} \forall t \in \mathcal{T}, \quad (4)$$

where the *decision of nature* at time epoch t is the health state TPM, Q_t, the *policy of nature* from epoch t onward is a vector of health state TPMs, denoted by $\boldsymbol{\theta}_t = (\underbrace{Q_t, Q_{t+1}, \ldots, Q_{T-1}}_{T-t})$, and the set $\Theta_t = \{(\underbrace{Q_t, Q_{t+1}, \ldots, Q_{T-1}}_{T-t}) | \, Q_k \in \mathcal{Q}_k, \forall k \in \{t, t+1, \ldots, T-1\}\}$, represents the set of all admissible *policies of nature* from epoch t onwards. In this robust context, the optimal value function can be interpreted as the maximum worst-case expected total discounted reward from epoch t until the patient reaches the absorbing state.

Based on Theorem 1 of Nilim and El Ghaoui (2005) as well as Theorem 2.2 of Iyengar (2005), when the uncertainty set, $\mathcal{Q}_t, \forall t \in \mathcal{T}\backslash\{T\}$, has the rectangular uncertainty property, the optimal solution of the RMDP-TM model can be obtained by a deterministic Markovian policy, and the optimal value function of the RMDP-TM can be written recursively as follows:

$$v_t^{\text{RMDP-TM}}(s_t) = \begin{cases} \max\limits_{\boldsymbol{\alpha}_t(s_t)\in\mathcal{A}_t} \Big\{ r_t(s_t, \boldsymbol{\alpha}_t(s_t)) + (1 - p_t^{\text{E}}(s_t, \boldsymbol{\alpha}_t(s_t)))\lambda \min\limits_{\boldsymbol{q}_{t,\ell_t}\in\mathcal{Q}_{t,\ell_t}} \sum\limits_{\ell_{t+1}\in\mathcal{L}} q_{t,\ell_t}(\ell_{t+1}) \\ \times v_{t+1}^{\text{RMDP-TM}}(\ell_{t+1}, \boldsymbol{m}_{t+1}(\boldsymbol{\alpha}_t(s_t))) \Big\}, & \forall s_t = (\ell_t, \boldsymbol{m}_t) \in \mathcal{L}\times\mathcal{M}, t \in \mathcal{T}\backslash\{T\}, \\ r_T(s_T), & \forall s_T = (\ell_T, \boldsymbol{m}_T) \in \mathcal{L}\times\mathcal{M}, \\ 0, & \forall s_t \in \mathcal{D}, \end{cases} \quad (5)$$

where the minimization problem presented in Equation (5) is often referred to as the *inner problem*:

$$\sigma_t(s_t, \boldsymbol{\alpha}_t(s_t)) = \min_{\boldsymbol{q}_{t,\ell_t}\in\mathcal{Q}_{t,\ell_t}} \sum_{\ell_{t+1}\in\mathcal{L}} q_{t,\ell_t}(\ell_{t+1}) v_{t+1}^{\text{RMDP-TM}}(\ell_{t+1}, \boldsymbol{m}_{t+1}(\boldsymbol{\alpha}_t(s_t)) \quad \forall t \in \mathcal{T}. \quad (6)$$

The RMDP-TM is a time-varying RMDP model since the decisions of nature are allowed to vary across time epochs. The time-invariant counterpart of the RMDP-TM, which requires the decision of nature to be the same for every time epoch, is a more difficult

RMDP model due to the correlation of the decisions of nature across the entire time horizon. The optimal value function of the time-invariant counterpart of the RMDP-TM can be written as follows:

$$v_t^{\text{TI-RMDP-TM}}(s_t) = \begin{cases} \max\limits_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} \min\limits_{\boldsymbol{\theta}_t^{\text{s}}\in\Theta_t} \mathbb{E}_{s_t}^{\boldsymbol{\pi},\boldsymbol{\theta}_t^{\text{s}}} \left\{ \sum\limits_{k=t}^{T-1} \lambda^{k-t} r_k(s_k, \boldsymbol{\alpha}_k(s_k)) + \lambda^{T-t} r_T(s_T) \right\}, \forall\, s_t \in \mathcal{L}\times\mathcal{M}, \\ 0, \qquad\qquad \text{otherwise}, \end{cases} \tag{7}$$

$\forall t \in \mathcal{T}$, where $\boldsymbol{\theta}_t^{\text{s}} \triangleq (\underbrace{Q, Q, \ldots, Q}_{T-t})$ represents a *stationary policy of nature* from epoch t onward. In Section 3 we will provide sufficient conditions under which problems (5) and (7) are equivalent, i.e. when solving (5) generates a stationary policy of nature.

In addition to uncertainty in the health state TPMs, the probability of entering the absorbing state may also be subject to uncertainty. Frequently these transition probabilities are estimated using published statistical risk models. For some diseases there are multiple published risk models, possibly resulting in a range of estimates. Taking this uncertainty into account, the optimality equations (4) become:

$$v_t'(s_t) = \begin{cases} \max\limits_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} \min\limits_{(\boldsymbol{p}_t^{\text{E}},\theta_t)\in(\boldsymbol{\mathcal{P}}_t^{\text{E}},\Theta_t)} \mathbb{E}_{s_t}^{(\boldsymbol{\pi},\boldsymbol{p}_t^{\text{E}},\theta_t)} \left[\sum\limits_{k=t}^{T-1} \lambda^{k-t} r_k(s_k, \boldsymbol{\alpha}_k(s_k)) + \lambda^{T-t} r_T(s_T) \right], \forall\, s_t \in \mathcal{L}\times\mathcal{M}, \\ 0, \qquad\qquad \text{otherwise}, \end{cases}$$

$\forall t \in \mathcal{T}$, where

$$\boldsymbol{p}_t^{\text{E}} = (p_t^{\text{E}}(s_t, \boldsymbol{\alpha}_t(s_t)), p_{t+1}^{\text{E}}(s_{t+1}, \boldsymbol{\alpha}_{t+1}(s_{t+1})), \ldots, p_{T-1}^{\text{E}}(s_{T-1}, \boldsymbol{\alpha}_{T-1}(s_{T-1}))) \in \boldsymbol{\mathcal{P}}_t^{\text{E}}$$

is a vector of transition probabilities of entering the absorbing state from epoch t to $T-1$, and $\boldsymbol{\mathcal{P}}_t^{\text{E}}$ is a set containing all possible vectors of probabilities of entering the absorbing state from epoch t to $T-1$. Under the assumption that uncertainty about the ideal choice of risk models for estimating the transition probabilities of entering the absorbing state is at the discretion of the decision maker, solving for $v_t'(s_t)$ is equivalent to solving for $v_t^{\text{RMDP-TM}}(s_t)$ when $p_k^{\text{E}}(s_k, \boldsymbol{\alpha}_t(s_k))$ takes the maximum value for all $k = t, t+1, \ldots, T-1$. Thus, in the remainder of this article, we focus on the more challenging RMDP-TM problem in (4).

2.3. Uncertainty Set Formulation: Interval Model With Uncertainty Budget

The choice of the uncertainty set $\mathcal{Q}_t, \forall t \in \mathcal{T}$ plays an important role in determining the computational tractability of solving the RMDP-TM (4), and determining the conservativeness of the corresponding optimal solutions. Some authors have proposed elliptical

uncertainty set (Nilim and El Ghaoui 2005, Bertsimas et al. 2011). These provide a good representation of the confidence region of the TPM but they also lead to a nonlinear formulation of the inner problem (6) which can cause the RMDP to be difficult to solve in practical use.

The interval matrix (IM) model presented in Nilim and El Ghaoui (2005) is perhaps one of the simplest uncertainty set formulations that also possesses the rectangular uncertainty property. In the IM model, $\mathcal{Q}_t^{\text{IM}}$, the uncertainty set of the TPM is a Cartesian product of the uncertainty set for each row ℓ_t of the TPM, which can be written as $\mathcal{Q}_t^{\text{IM}} = \prod_{\ell_t \in \mathcal{L}} \mathcal{Q}_{t,\ell_t}^{\text{IM}}$, where $\mathcal{Q}_{t,\ell_t}^{\text{IM}}$ is the uncertainty set for row ℓ_t of matrix Q_t with the following form:

$$\mathcal{Q}_{t,\ell_t}^{\text{IM}} = \left\{ \boldsymbol{q}_{t,\ell_t} \in \mathbb{R}_+^{|\mathcal{L}|} : \sum_{\ell_{t+1} \in \mathcal{L}} q_{t,\ell_t}(\ell_{t+1}) = 1,\ q_{t,\ell_t}^l(\ell_{t+1}) \le q_{t,\ell_t}(\ell_{t+1}) \le q_{t,\ell_t}^u(\ell_{t+1}),\ \forall \ell_{t+1} \in \mathcal{L} \right\},$$

$\forall \ell_t \in \mathcal{L}$, $\forall t \in \mathcal{T} \backslash \{T\}$. The quantities $q_{t,\ell_t}^l(\ell_{t+1})$ and $q_{t,\ell_t}^u(\ell_{t+1})$ denote the lower and upper bounds of $q_{t,\ell_t}(\ell_{t+1})$, respectively. Many approaches have been developed to generate simultaneous confidence intervals for multinomial proportions from observational datasets. We refer the interested readers to Gold (1963), Goodman (1965), Fitzpatrick and Scott (1987) and Sison and Glaz (1995) for different calculation methods.

To mitigate the conservativeness of the IM model while maintaining the rectangular uncertainty property, we combine the IM model with an additional *uncertainty budget* constraint to control the size of the uncertainty set. The uncertainty budget parameter was originally proposed in the math programming literature (Bertsimas and Sim 2004, Bertsimas et al. 2011) as a means to tradeoff the protection level of the constraints and the degree of conservatism of the solution. This motivated us to use a similar approach to see if there are benefits to limiting variation in transition probabilities within an uncertainty set to influence how conservative the (worst case) solutions are for the traditional interval model. Other methods have also been proposed to address the conservativeness of the solutions in the context of RMDP. For example, Mannor et al. (2016) presented a "Lighting does not strike twice" model in which the number of states whose parameters can deviate from their nominal values is bounded by a given budget. Kaufman et al. (2011) presented an infinite-horizon RMDP for optimizing the time to undergo a living-donor liver transplantation where relative entropy bounds proposed by Nilim and El Ghaoui (2005) were used to construct the uncertainty set for transition probabilities and the size of the uncertainty set were controlled by the confidence level.

We refer to our model as the interval model with uncertainty budget (IMUB). For any $\ell_t, \ell_{t+1} \in \mathcal{L}$, we define the maximal left/right-hand-side variation associated with probability $q_{t,\ell_t}(\ell_{t+1})$ as the absolute difference between the lower/upper bound of the transition probability, $q_{t,\ell_t}(\ell_{t+1})$, and its nominal value, $\hat{q}_{t,\ell_t}(\ell_{t+1})$; and denote them by $\delta^l_{t,\ell_t}(\ell_{t+1}) \triangleq \hat{q}_{t,\ell_t}(\ell_{t+1}) - q^l_{t,\ell_t}(\ell_{t+1})$ and $\delta^u_{t,\ell_t}(\ell_{t+1}) \triangleq q^u_{t,\ell_t}(\ell_{t+1}) - \hat{q}_{t,\ell_t}(\ell_{t+1})$, respectively. In addition, we define the degree of left/right-hand-side variation as the proportion of variation from the lower/upper bound of $q_{t,\ell_t}(\ell_{t+1})$ to its nominal value, and denote them by $z^l_{t,\ell_t}(\ell_{t+1})$ and $z^u_{t,\ell_t}(\ell_{t+1})$, respectively. Therefore, $\forall \ell_t \in \mathcal{L}, t \in \mathcal{T}\backslash\{T\}$, if $q_{t,\ell_t}(\ell_{t+1}) \le \hat{q}_{t,\ell_t}(\ell_{t+1})$, then $z^l_{t,\ell_t}(\ell_{t+1}) = (\hat{q}_{t,\ell_t}(\ell_{t+1}) - q_{t,\ell_t}(\ell_{t+1}))/\delta^l_{t,\ell_t}(\ell_{t+1})$ and $z^u_{t,\ell_t}(\ell_{t+1}) = 0$; otherwise, $z^u_{t,\ell_t}(\ell_{t+1}) = (q_{t,\ell_t}(\ell_{t+1}) - \hat{q}_{t,\ell_t}(\ell_{t+1}))/\delta^u_{t,\ell_t}(\ell_{t+1})$ and $z^l_{t,\ell_t}(\ell_{t+1}) = 0$. We assume there is no variation when $\hat{q}_{t,\ell_t}(\ell_{t+1}) = 0$. i.e. $z^l_{t,\ell_t}(\ell_{t+1})$ and $z^u_{t,\ell_t}(\ell_{t+1})$ equal 0.

The uncertainty budget on row ℓ_t of the TPM, Q_t, is denoted by Γ_{t,ℓ_t}. It defines a limit on the total allowable variations of the probabilities from their nominal values, measured by the sum of $z^l_{t,\ell_t}(\ell_{t+1})$ and $z^u_{t,\ell_t}(\ell_{t+1})$ for all ℓ_{t+1}. Thus Γ_{t,ℓ_t} is limited to the range from 0 to $|\mathcal{L}|$. The complete IMUB for row ℓ_t of the TPM, Q_t, can be written as follows:

$$\mathcal{Q}^{\text{IMUB}}_{t,\ell_t}(\Gamma_{t,\ell_t}) = \left\{ \boldsymbol{q}_{t,\ell_t} \in \mathbb{R}^{|\mathcal{L}|}_+ \;\middle|\; \begin{array}{ll} q_{t,\ell_t}(\ell_{t+1}) = \hat{q}_{t,\ell}(\ell_{t+1}) - \delta^l_{t,\ell_t}(\ell_{t+1}) z^l_{t,\ell_t}(\ell_{t+1}) & \\ \qquad\qquad + \delta^u_{t,\ell_t}(\ell_{t+1}) z^u_{t,\ell_t}(\ell_{t+1}), & \forall \ell_{t+1} \in \mathcal{L}, \\ \sum\limits_{\ell_{t+1}\in\mathcal{L}} q_{t,\ell_t}(\ell_{t+1}) = 1, & \\ \sum\limits_{\ell_{t+1}\in\mathcal{L}} \left(z^l_{t,\ell_t}(\ell_{t+1}) + z^u_{t,\ell_t}(\ell_{t+1}) \right) \le \Gamma_{t,\ell_t}. & \\ z^l_{t,\ell_t}(\ell_{t+1}) \cdot z^u_{t,\ell_t}(\ell_{t+1}) = 0, & \forall \ell_{t+1} \in \mathcal{L}, \\ 0 \le z^l_{t,\ell_t}(\ell_{t+1}), z^u_{t,\ell_t}(\ell_{t+1}) \le 1, & \forall \ell_{t+1} \in \mathcal{L} \\ 0 \le q_{t,\ell_t}(\ell_{t+1}) \le 1, & \forall \ell_{t+1} \in \mathcal{L}, \end{array} \right\} \tag{8}$$

The first equation in (8) represents the variable $q_{t,\ell_t}(\ell_{t+1})$ in terms of its nominal value, and the degree of left-hand-side and right-hand-side variation. Since $q_{t,\ell_t}(\ell_{t+1})$ can only be on one side of its nominal value, $z^l_{t,\ell_t}(\ell_{t+1})$ and $z^u_{t,\ell_t}(\ell_{t+1})$ cannot be positive simultaneously. Therefore, the fourth constraint guarantees that either $z^l_{t,\ell_t}(\ell_{t+1})$ or $z^u_{t,\ell_t}(\ell_{t+1})$ or both are zero. The second and sixth constraints guarantee that the variables, $q_{t,\ell_t}(\ell_{t+1})$, $\forall \ell_{t+1} \in \mathcal{L}$, are confined to the probability simplex. The fifth constraint provides the lower and upper bounds for the variables $z^l_{t,\ell_t}(\ell_{t+1})$ and $z^u_{t,\ell_t}(\ell_{t+1})$. The third constraint is the uncertainty budget constraint, which requires the total degree of uncertainty on row ℓ_t of the TPM,

Q_t, to be less than or equal to Γ_{t,ℓ_t}. The IMUB model of the entire TPM, Q_t, $\forall t \in \mathcal{T}$, can be written as $\mathcal{Q}_t^{\text{IMUB}}(\Gamma_{t,\ell_t}) = \prod_{\ell_t \in \mathcal{L}} \mathcal{Q}_{t,\ell_t}^{\text{IMUB}}(\Gamma_{t,\ell_t})$, $\forall\, t \in \mathcal{T}$.

Remark 1: Notice that when $\Gamma_{t,\ell_t} = 0$, $\forall \ell_t \in \mathcal{L}, t \in \mathcal{T}$, we have $q_{t,\ell_t}(\ell_{t+1}) = \hat{q}_{t,\ell_t}(\ell_{t+1})$, $\forall \ell_t, \ell_{t+1} \in \mathcal{L}, t \in \mathcal{T}$, and the RMDP-TM is the MDP model. Likewise, when $\Gamma_{t,\ell_t} = |\mathcal{L}|$, $\forall \ell_t \in \mathcal{L}, t \in \mathcal{T}$, the uncertainty set corresponds to the IM model. Therefore, the uncertainty budget controls the conservativeness of the optimal policy of the RMDP-TM relative to the MDP model.

3. Analysis and Algorithm for RMDP-TM

In this section, we present theoretical analysis of the RMDP-TM that can be used to establish computationally efficient methods for solving the RMDP-TM. We also provide sufficient conditions under which the time-invariant counterpart of the RMDP-TM can be solved. Complete proofs of Propositions 1–3 are available in Appendix A.

In light of the recursive structure in Equation (5), the RMDP-TM can be solved using the robust dynamic programming (RDP) algorithm proposed by Nilim and El Ghaoui (2005). In the RDP algorithm, the expected total discounted worst-case value-to-go is calculated by solving the inner problem (6) at each iteration. Since the inner problem needs to be solved $|\mathcal{S}| \cdot |\mathcal{T} \backslash \{T\}|$ times, the extra computational cost of solving the RMDP-TM depends on how efficiently the inner problem can be solved.

The inner problem (6) with IMUB model is the following nonlinear program:

$$
\begin{aligned}
\min \quad & \sigma_t^{\text{IMUB-NLP}}(s_t, \boldsymbol{\alpha}_t(s_t), \Gamma_{t,\ell_t}) \\
& \qquad = \sum_{\ell_{t+1}\in\mathcal{L}} q_{t,\ell_t}(\ell_{t+1}) v_{t+1}^{\text{RMDP-TM}}(\ell_{t+1}, \boldsymbol{m}_{t+1}(\boldsymbol{\alpha}_t(s_t))) \\
\text{s.t.} \quad & q_{t,\ell_t}(\ell_{t+1}) = \hat{q}_{t,\ell}(\ell_{t+1}) - \delta_{t,\ell}^l(\ell_{t+1}) z_{t,\ell}^l(\ell_{t+1}) \qquad (9) \\
& \qquad + \delta_{t,\ell}^u(\ell_{t+1}) z_{t,\ell}^u(\ell_{t+1}), \forall \ell_{t+1} \in \mathcal{L}, \\
& \sum_{\ell_{t+1}\in\mathcal{L}} q_{t,\ell_t}(\ell_{t+1}) = 1, \qquad (10) \\
& \sum_{\ell_{t+1}\in\mathcal{L}} \left(z_{t,\ell_t}^l(\ell_{t+1}) + z_{t,\ell_t}^u(\ell_{t+1}) \right) \le \Gamma_{t,\ell}, \qquad (11) \\
& z_{t,\ell_t}^l(\ell_{t+1}) \cdot z_{t,\ell_t}^u(\ell_{t+1}) = 0, \forall \ell_{t+1} \in \mathcal{L}, \qquad (12) \\
& 0 \le z_{t,\ell_t}^l(\ell_{t+1}), z_{t,\ell_t}^u(\ell_{t+1}) \le 1, \forall \ell_{t+1} \in \mathcal{L}. \qquad (13) \\
& 0 \le q_{t,\ell_t}(\ell_{t+1}) \le 1, \forall \ell_{t+1} \in \mathcal{L}, \qquad (14)
\end{aligned}
$$

(IMUB-NLP)

The following proposition provides a linear reformulation of IMUB-NLP that is easier to solve.

PROPOSITION 1. *The IMUB-NLP is equivalent to the following IMUB-LP.*

$$
\begin{aligned}
\min \quad & \sigma_t^{\text{IMUB-LP}}(\ell_t, \boldsymbol{m}_t, \boldsymbol{\alpha}_t(s_t), \Gamma_{t,\ell}) \\
& \qquad = \sum_{\ell_{t+1}\in\mathcal{L}} q_{t,\ell_t}(\ell_{t+1}) v_{t+1}^{\text{RMDP-TM}}(\ell_{t+1}, \boldsymbol{m}_{t+1}(\boldsymbol{\alpha}_t(s_t))) \\
\text{s.t.} \quad & q_{t,\ell_t}(\ell_{t+1}) = \hat{q}_{t,\ell}(\ell_{t+1}) - \delta_{t,\ell}^l(\ell_{t+1}) z_{t,\ell}^l(\ell_{t+1}) \qquad (15) \\
& \qquad + \delta_{t,\ell}^u(\ell_{t+1}) z_{t,\ell}^u(\ell_{t+1}), \forall \ell_{t+1} \in \mathcal{L}, \\
& \sum_{\ell_{t+1}\in\mathcal{L}} q_{t,\ell_t}(\ell_{t+1}) = 1, \qquad (16) \\
& \sum_{\ell_{t+1}\in\mathcal{L}} \left(z_{t,\ell_t}^l(\ell_{t+1}) + z_{t,\ell_t}^u(\ell_{t+1}) \right) \le \Gamma_{t,\ell}, \qquad (17) \\
& 0 \le z_{t,\ell_t}^l(\ell_{t+1}), z_{t,\ell_t}^u(\ell_{t+1}) \le 1, \forall \ell_{t+1} \in \mathcal{L}. \qquad (18) \\
& 0 \le q_{t,\ell_t}(\ell_{t+1}) \le 1, \forall \ell_{t+1} \in \mathcal{L}, \qquad (19)
\end{aligned}
$$

(IMUB-LP)

Proposition 1 states that the nonlinear constraint in the IMUB-NLP can be removed without changing the optimal value of the objective function. The formal proof of Propo-

sition 1 is given in the Appendix A. The basic idea of the proof is to show that for a given feasible solution of IMUB-LP, a feasible solution to IMUB-NLP can be constructed. On the other hand, it is clear that any feasible solution to IMUB-NLP is feasible to IMUB-LP. Therefore, the two problems are equivalent.

Algorithm 1 A fast algorithm to solve the inner problem (6) with the IM model

for $i = 1 \to |\mathcal{L}|$ **do**
  set $y_i^l \leftarrow q_{t,\ell_t}^l(\ell(i))$, $y_i^u \leftarrow q_{t,\ell_t}^u \ell(i)$, $c_i \leftarrow v_{t+1}^{\text{RMDP-TM}}(\ell(i), \boldsymbol{m}_{t+1}(\boldsymbol{\alpha}_t(s_t)))$
end for
for $\tau = 1 \to |\mathcal{L}|$ **do**
  set $\beta \leftarrow c_\tau$ and $\sigma \leftarrow (1 - \sum_{i=1}^{|\mathcal{L}|} y_i^l)\beta + \sum_{i=1}^{|\mathcal{L}|}(y_i^u - y_i^l)\min\{c_i - \beta, 0\} + \sum_{i=1}^{|\mathcal{L}|} c_i y_i^l$
  if $(\tau == 1)$ or $(\tau > 1$ and $\sigma > \sigma_t^{\text{IM}}(s_t, \boldsymbol{\alpha}_t(s_t)))$ **then**
    $\sigma_t^{\text{IM}}(s_t, \boldsymbol{\alpha}_t(s_t)) \leftarrow \sigma$
  else
    Next τ
  end if
end for
Return $\sigma_t^{\text{IM}}(s_t, \boldsymbol{\alpha}_t(s_t))$

Next, we show that when the uncertainty set corresponds to the IM model, the inner problem (6) can be solved by Algorithm 1 has complexity $O(|\mathcal{L}|)$.

PROPOSITION 2. *Algorithm (1) generates an optimal solution to the inner problem (6) with the IM model, and has complexity,* $O(|\mathcal{L}|)$.

The formal proof of Proposition 2 is in the Appendix A. The basic idea is to construct the dual problem and prove that the optimal objective function value could be obtained by Algorithm (1).

The time-invariant counterpart of the RMDP-TM model is not solvable by the RDP algorithm proposed by Nilim and El Ghaoui (2005) due to the dependency of the nature's decisions across time epochs. This problem could be solved by using a large semidefinite program proposed by Wiesemann et al. (2013), which does not exploit the recursive structure of the optimal value function; however, this is much more computationally intensive

than the RDP algorithm and may not be feasible for large practical problems. Nilim and El Ghaoui (2005) find that the gap between the time-invariant RMDP and the time-varying RMDP goes to zero when the time horizon goes to infinity. For the RMDP-TM, we find sufficient conditions under which the RMDP-TM equals its time-invariant counterpart for a finite horizon. We begin with a definition of nonincreasing worst-case (NIWC) TPM that is relevant to the proposition.

DEFINITION 1. The TPM, $Q_t^{\text{NIWC},\Gamma}$, is called an *NIWC TPM* in the uncertainty set $\mathcal{Q}_t^{\text{IMUB}}(\Gamma) = \prod_{\ell_t \in \mathcal{L}} \mathcal{Q}_{t,\ell_t}^{\text{IMUB}}(\Gamma)$, if $\forall\, \ell_t \in \mathcal{L}$, row ℓ_t of the TPM, $Q_t^{\text{NIWC},\Gamma}$, denoted by $\boldsymbol{q}_{t,\ell_t}^{\text{NIWC},\Gamma}$, is the optimal solution of the following problem:

$$\beta_t(\ell_t) = \min_{\boldsymbol{q}_{t,\ell_t} \in \mathcal{Q}_{t,\ell_t}^{\text{IMUB}}(\Gamma)} \sum_{\ell_{t+1} \in \mathcal{L}} q_{t,\ell_t}(\ell_{t+1}) c(\ell_{t+1}) \quad \forall t \in \mathcal{T}\backslash\{T\}. \tag{20}$$

where the coefficients of the objective function, $\{c(\ell_{t+1})\}_{\ell_{t+1}\in\mathcal{L}}$, are nonincreasing in the health state, ℓ_{t+1}.

Remark 2: Notice that when $c(\ell_{t+1})$ equals $v_{t+1}^{\text{RMDP-TM}}(\ell_{t+1}, \boldsymbol{m}_{t+1}(\boldsymbol{\alpha}_t(s_t)))$, $\forall \ell_{t+1} \in \mathcal{L}$, the problem (20) is equivalent to the inner problem (6).

PROPOSITION 3. *For the RMDP-TM with optimal value function shown in (5), if the following conditions hold:*

(I): The uncertainty set of the TPM, Q_t, is $\mathcal{Q}_t^{IM}$, an IM model, $\forall t \in \mathcal{T}\backslash\{T\}, \boldsymbol{m}_t \in \mathcal{M}$,

(II): $\mathcal{Q}_t^{IM} = \mathcal{Q}_{t'}^{IM}, \forall t, t' \in \mathcal{T}\backslash\{T\}$,

(III): $r_t(\ell_t, \boldsymbol{m}_t, \boldsymbol{\alpha}_t(s_t))$ is nonincreasing in ℓ_t, $\forall\, \boldsymbol{m}_t \in \mathcal{M}$, $\boldsymbol{\alpha}_t(s_t) \in \mathcal{A}_t$, and $t \in \mathcal{T}\backslash\{T\}$, and $r_T(\ell_T, \boldsymbol{m}_T)$ is nonincreasing in ℓ_T, $\forall\, \boldsymbol{m}_T \in \mathcal{M}$,

(IV): $p_t^E(\ell_t, \boldsymbol{m}_t, \boldsymbol{\alpha}_t(s_t))$ is nondecreasing in ℓ_t, $\forall\, \boldsymbol{m}_t \in \mathcal{M}$, $\boldsymbol{\alpha}_t(s_t) \in \mathcal{A}_t$, and $t \in \mathcal{T}\backslash\{T\}$, and

(V): $Q_{T-1}^{NIWC,|\mathcal{L}|}$, has the increasing failure rate (IFR) property, namely, $\sum_{\ell=\ell(k)}^{\ell(|\mathcal{L}|)} q_{t,\ell_t}^{NIWC}(\ell)$ is non-decreasing in $\ell_t \in \mathcal{L}$ for all $\ell(k) \in \mathcal{L}$ (Barlow and Proschan 1965).

then

(a): the optimal value function, $v_t^{RMDP\text{-}TM}(\ell_t, \boldsymbol{m}_t)$, of the RMDP-TM is nonincreasing in ℓ_t, $\forall\, \boldsymbol{m}_t \in \mathcal{M}, t \in \mathcal{T}\backslash\{T\}$, and

(b): the optimal policy of nature is $(Q_{T-1}^{NIWC,|\mathcal{L}|}, Q_{T-1}^{NIWC,|\mathcal{L}|}, \ldots, Q_{T-1}^{NIWC,|\mathcal{L}|})$, therefore, it is stationary.

The proof of Proposition 3 is in the Appendix A, and is done by induction. These sufficient conditions identify a special case when the time-invariant counterpart of the RMDP-TM can be solved. Conditions (I) and (II) state that the uncertainty sets of the health state TPM for all decision epochs are the same IM model. Condition (III) states that the immediate reward and the terminal reward are nonincreasing as the health state becomes worse for any treatment state. Condition (IV) states that the probability of entering the absorbing state is nondecreasing in health states for any treatment states. These assumptions are reasonable in making medical treatment decisions where a poorer health state is usually associated with a lower reward, and a higher probability of entering the absorbing state. In simple terms, the assumption of the IFR property in Condition (V) states that the worse a patient's health state the more likely it is to get worse. It has been widely used to analyze the structure of the optimal policy for many applications of MDPs and has been observed to hold empirically in many disease contexts (Mason et al. 2014, Shechter et al. 2008, Kurt et al. 2011, Alagoz et al. 2004, 2007). Proposition 3 establishes conditions under which the RMDP-TM (4) and its time-invariant counterpart (7) are equivalent. In addition, based on Proposition 3, the computational time for solving a special class of the RMDP-TMs, satisfy the conditions in Proposition 3, can be further reduced to the same as solving an MDP with the TPM set to be the NIWC TPM.

4. Case Study: Optimizing Treatment Decisions for Type 2 Diabetes

In this section, we present a case study illustrates the application of the proposed model to optimize the treatment decisions for glycemic control for patients with type 2 diabetes. This case study also serves to analyze the performance of the resulting optimal policies.

4.1. Background and RMDP-TM Formulation of the Glycemic Control Model

A central focus of managing type 2 diabetes is glycemic control which involves the regulation of blood glucose levels over time. Glycemic control aims to avoid acute daily symptoms of hyperglycemia, to avoid instability in blood glucose over time, and to prevent or delay the development of diabetes-related complications associated with the high blood glucose levels. Glycated hemoglobin (HbA1c) is a commonly used measure of average blood glucose concentration over time. A high HbA1c indicates poor glycemic control, and the need to initiate medication(s) such as one of several oral medications or insulin. HbA1c is obtained via a simple blood test at a recommended frequency of every three months (American Diabetes Association 2017).

Zhang et al. (2014) presented and validated a population-based glycemic control model based on a finite horizon Markov chain. This model was used to compare the health benefit, i.e. the quality-adjusted life-years (QALYs) gained, and the total medication costs of different treatment regimens for individuals newly diagnosed with type 2 diabetes.The key assumptions of the model include: (1) the HbA1c state transition probabilities are independent of the medication being used except for insulin whose dose can be adjusted automatically via an infusion system to maintain patient's blood sugar; (2) medication results in a proportional decrease in HbA1c based on the medication being used based on empirical estimates using observational data for patients taking each medication; and (3) other metabolic measures, e.g., blood pressure, cholesterol, which may affect the risk of having complications are assumed to be maintained at clinically recommended levels by other drugs and are not included in the model.

We used the glycemic control model presented in Zhang et al. (2014) to create an RMDP-TM to optimize the treatment decisions for glycemic control for patients with type 2 diabetes in light of uncertainty in HbA1c transition probability estimates caused by statistical variations. Following is a description of the model in this context.

Time horizon. The time horizon starts from the age at diagnosis, e.g., 55 years old for females and 53 years old for males in the United States (Centers for Disease Control and Prevention 2013), and spans the remaining patient's lifetime. The time horizon is discretized into three-month intervals based on the recommended frequency of performing the HbA1c test (American Diabetes Association 2017)

States. The health states in the model include 10 HbA1c states, $\mathcal{L} = \{\ell(1), \ell(2), \ldots, \ell(10)\}$, defined by a discrete set of clinically relevant ranges of HbA1c levels. As a function of age, the mean HbA1c value for each HbA1c state increases linearly, reflecting the expected deterioration of glycemic control as the patient ages (Sinha et al. 2010). The ranges of HbA1c levels used for categorizing HbA1c states and the mean HbA1c levels for HbA1c states are shown in Appendix C Tables 1 and 2. We consider three hyperglycemia-lowering medications: metformin, sulfonylurea, and insulin because they were shown to be the most cost-effective in Zhang et al. (2014). The treatment states are represented as $3-$tuple binary vectors in which the first, the second, and the third elements represent the usage status for metformin, sulfonylurea, and insulin, respectively. The absorbing state, $\mathcal{D}$, includes major diabetes-related complications: fatal or non fatal macrovascular events (such as ischemic

heart disease, congestive heart failure, and stroke); fatal or non fatal microvascular events (such as blindness, renal failure, and limb amputation); and severe hypoglycemia requiring hospitalization; and death from other causes.

Actions. The action is the selection of which medication(s) to initiate at each time epoch. Treatment results in a proportional decrease in HbA1c according to the medication effects estimated from a longitudinal administrative claims dataset including HbA1c records and pharmacy claims (shown in Zhang et al. (2014)). There are no further treatment changes once insulin is initiated, as it is assumed to maintain control of the HbA1c level.

Rewards. The immediate reward represents a patient's QALYs between time epochs. QALYs are widely used for evaluating the health outcome for treatment and health policy decisions (Gold et al. 2002). QALYs adjust a year of life proportionally based on *utilities* that represent the quality of health that the patient experiences. In this case study, each year is assigned a value between 0 (death) and 1 (perfect health) where the exact value depends on the occurrence of side effects of medication and the disutility of hyperglycemia symptom. For $\forall\, s_t \in \mathcal{L} \times \mathcal{M}$, the immediate reward is calculated as follows:

$$r_t(s_t, \boldsymbol{\alpha}_t(s_t)) = 0.25 \times (1 - D^{\text{hyper}}(\boldsymbol{\alpha}_t(s_t))(1 - D^{\text{med}}(\boldsymbol{\alpha}_t(s_t)), \forall \ell_t \in \mathcal{L},$$

where $D^{\text{hyper}}(\boldsymbol{\alpha}_t(s_t))$ represents the disutility of hyperglycemia symptom when HbA1c level is above 8%, $D^{\text{med}}(\boldsymbol{\alpha}_t(s_t))$ represents the disutility of taking medications (a quantity that measures the side effects of medications) during the 3-month period as shown in Table 1 of Zhang et al. (2014). For $s_t = \mathcal{D}$, $r_t(s_t, \boldsymbol{\alpha}_t(s_t))$ is set to be zero. The terminal rewards, $r_T(s_T)$, are set to be 2.24 years for females and 2.05 years for males based on U.S. life tables (Arias 2011).

Probabilities. The probability of entering the absorbing state is calculated as follows: $p_t^{\mathcal{D}}(s_t, \boldsymbol{\alpha}_t(s_t)) = p_t^{\text{O}} + p_t^{\text{macro}}(\ell_t, \boldsymbol{\alpha}_t(s_t)) + p_t^{\text{micro}}(\ell_t, \boldsymbol{\alpha}_t(s_t)) + p_t^{\text{hypo}}(\boldsymbol{\alpha}_t(s_t))$. The probabilities of macro- and micro-vascular events, $p_t^{\text{macro}}(\ell_t, \boldsymbol{\alpha}_t(s_t))$ and $p_t^{\text{micro}}(\ell_t, \boldsymbol{\alpha}_t(s_t))$ were estimated by using the United Kingdom Prospective Diabetes Study (UKPDS) outcome models in the forms of survival functions (Clarke et al. 2004). The probabilities of severe hypoglycemia, $p_t^{\text{hypo}}(\boldsymbol{\alpha}_t(s_t))$, were also obtained from UKPDS Group (1995). The probabilities of death from other causes were obtained from mortality tables from the Centers for Disease Control and Prevention (Centers for Disease Control and Prevention 2012).

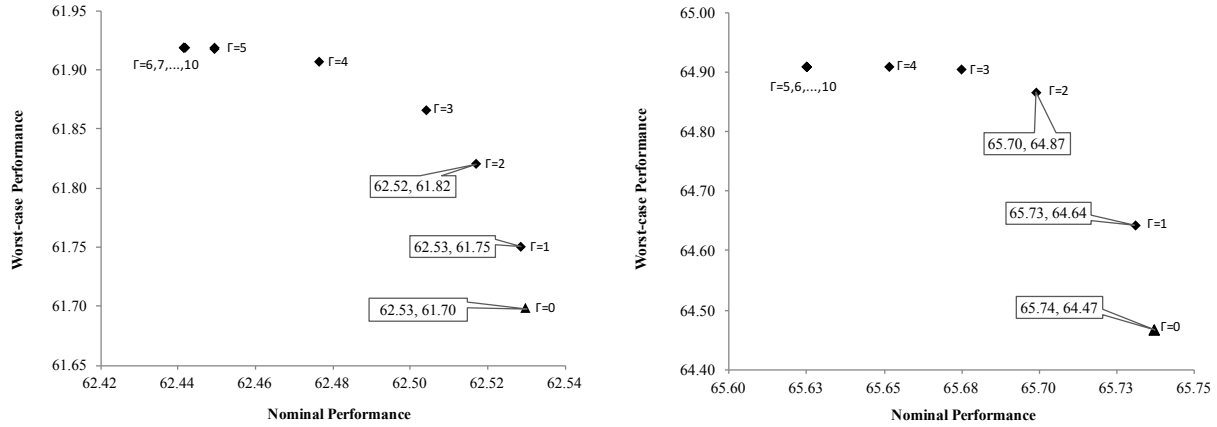
A retrospective administrative claims dataset with linked laboratory data from a large, nationwide US health plan was used to obtain the MLE and the left-hand-side and right-hand-side variations of HbA1c transition probabilities Zhang et al. (2014). The population meeting criteria for our study (337 males and 272 females) were 1) aged 40 years or older; 2) have been diagnosed with type 2 diabetes between 1995 and 2010; 3) have received the first non-insulin glucose-lowering medication at least 6 months after enrollment; and 4) have accumulated 15 HbA1c records within 5 years of continuous enrollment, along with complete pharmacy claims records.

The nominal HbA1c TPMs were estimated using maximum likelihood estimation based on the proportion of transitions between health states. As mentioned in Section 2.3, there are multiple ways to estimate the confidence region for each row of the HbA1c TPM. We used the method proposed in Sison and Glaz (1995) based on truncated Poisson approximation to calculate the 99% confidence intervals for HbA1c transition probabilities. The calculation was done by using the "Multinomial CI" R package available at `https://cran.r-project.org/web/packages/MultinomialCI/index.html`. The nominal HbA1c transition probabilities and the corresponding left-hand-side and right-hand-side variations are shown in Appendix B Tables 1–6.

The goal of the RMDP-TM version of the glycemic control model is to maximize the worst-case expected total QALYs from the time of diagnosis to the development of the first diabetes-related complication or death. Given the initial HbA1c state distribution, $\mathbf{\Pi} = (\Pi_1, \ldots, \Pi_{10})$ shown in Appendix Tables 1 and 2, the QALY results presented in Section 4.2 represent the worst-case expected total QALYs from birth to the first diabetes-related complication or death, and are calculated as $v^{\text{RMDP-TM}^*} = \text{LYs}_0 + \sum_{i=1}^{10} \Pi_i v_1^{\text{RMDP-TM}}(\ell(i), (0,0,0))$ where LYs_0 represents the expected life years from birth to the time of diagnosis, and the vector, $(0,0,0)$, represents the initial treatment state for patients not on any diabetes medication at the time of diagnosis.

4.2. Treatment Policy Performance Comparison

To compare the performance of various treatment policies, we define the *nominal performance* of a policy to be the expected total discounted reward under the nominal TPMs; and the *worst-case performance* of a policy to be the expected total discounted reward under the worst-case criterion when the TPM can vary over its uncertainty set. Theoretical



(a) Males (b) Females

Figure 1 **The theoretical nominal performance versus the theoretical worst-case performance based on quality-adjusted life-years to the first major complication for the MDP-optimal policy (i.e. the RMDP-TM-optimal policy with $\Gamma = 0$) (solid triangle), and RMDP-TM optimal policies with $\Gamma > 0$ (solid diamond).**

nominal and worst-case performance are presented in Figures 1a and 1b for males and females, respectively.

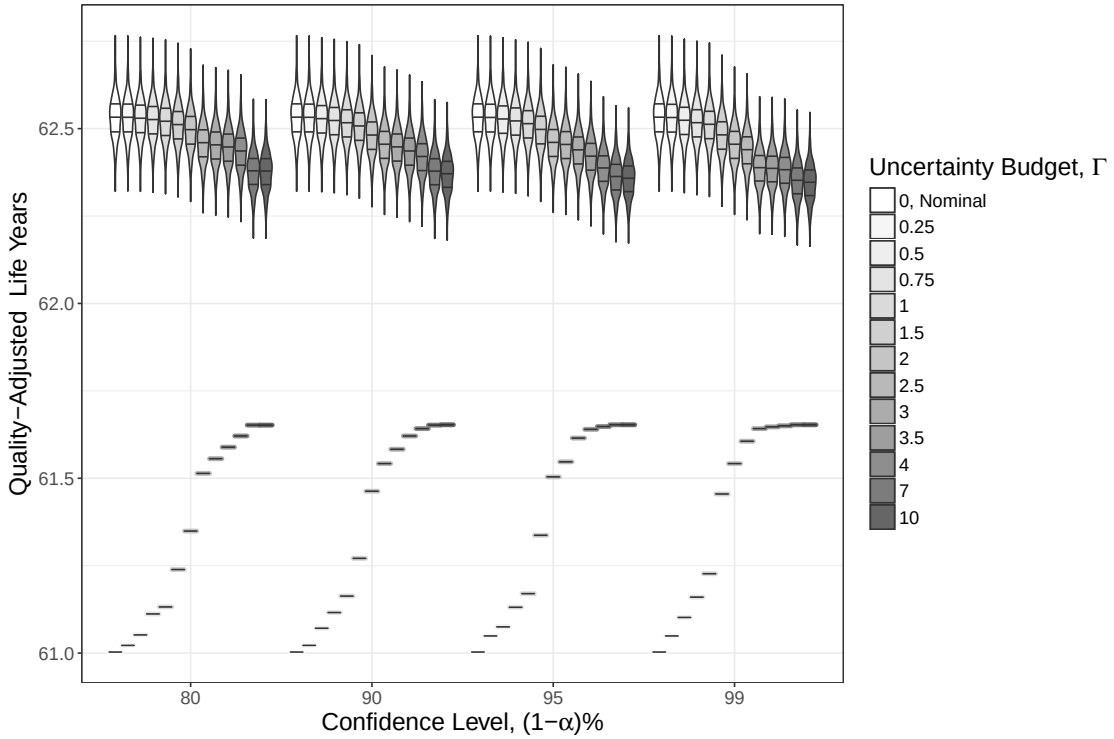
The MDP-optimal policy (i.e. the RMDP-TM-optimal policy with $\Gamma = 0$) results in the highest theoretical nominal performance, the expected QALYs are 62.53 for males and 65.74 for females. However, its corresponding theoretical worst-case performance is the lowest: 61.70 QALYs for males (a reduction of 0.82 QALYs from its nominal performance) and 64.47 QALYs for females (a reduction of 1.27 QALYs from its nominal performance). To put this in perspective, these reductions are more than an order of magnitude greater than the use of aspirin for secondary prevention of myocardial infarction in 45-year-old men, an important intervention, which has been estimated to provide a QALY gain of 0.04 per patient (Pignone et al. 2006). On the other hand, the IM verison of the RMDP-TM (i.e., $\Gamma = 10$) is very conservative because it results in the lowest theoretical nominal performance: 62.44 QALYs for males (a reduction of 0.09 QALYs from the theoretical nominal performance of the MDP-optimal policy) and 65.63 QALYs for females (a reduction of 0.11 QALYs from the theoretical nominal performance of the MDP-optimal policy). As the Γ value changes, we see a significant tradeoff between the value functions for the nominal and worst-case performance.

We also conducted a Monte-Carlo simulation study to compare the simulated performance of all the policies. Some challenges exist for sampling transition probabilities as each row sums up to 1 and each element needs to be lie in the interval $[0, 1]$. We used

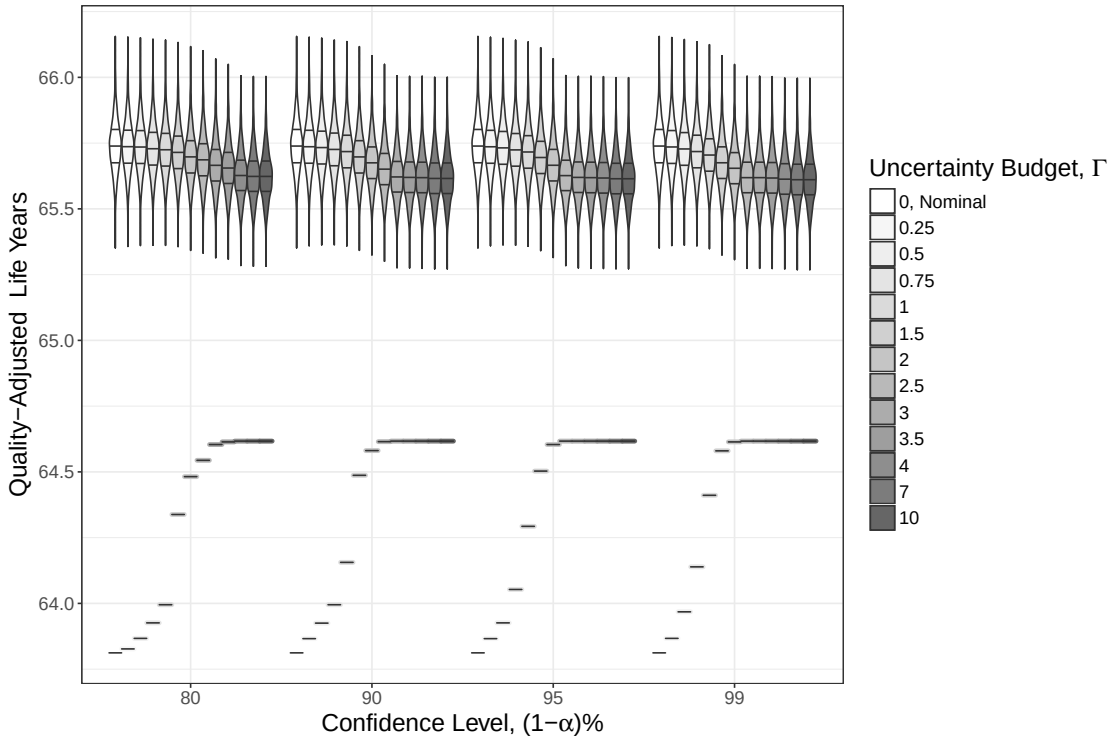
the sampling approach presented in Zhang et al. (2016) to sample transition probabilities according to the truncated multivariate normal distribution. For each experiment we sampled 1000 transition probability matrices. As motivated by Proposition 3, and the goal of evaluating the time invariant solution to RMDP-TM, we sampled the matrix once for each instance and used the sampled matrix in all decision epochs. Using this simulation approach we estimated various performance measures including (1) the sample mean of the value function by applying the MDP-optimal policy and the RMDP-TM optimal policies with Γ varying from 1 to 10; (2) the sampled distributions of the value function for the various policies; (3) the sampled worst-case observed over the 1000 samples; and (4) the standard deviation of the value function estimates over the 1000 samples. The base case results are for $\alpha = 0.01$, corresponding to 99% confidence interval. We also report results in which we varied the confidence level, $1-\alpha$, and the uncertainty budget, Γ, to evaluate the potential for using these parameters to control the performance characteristics of the policies generated using the RMDP-TM.

Figures 2a and 2b show *violin plots* include the sampled performance of RMDP-TM-optimal policies for varying choices of Γ in the range 0 to 10. Results at the top of the figure for the respective policies show the mean, the interquartile range and minimum and maximum estimates of the value function (i.e., the expected QALYs). Results at the bottom of the figure are for the theoretical worst-case performance for various choices of Γ values and confidence levels. A number of important insights can be gained from these results. First, the results are only moderately sensitive to the choice of confidence level, $1-\alpha$. Second, the choice of Γ significantly influences the worst-case performance of the policies and provides a means to tradeoff between the nominal performance and the worst-case performance estimated via sampling. Finally, and perhaps most importantly, there is a significant gap between the sampled worst-case expected QALYs and the theoretical worst-case expected QALYs, suggesting that a criteria focused on protecting against the theoretical worst-case, which is often suggested in prior literature, may be overly conservative for some decision makers.

Figures 3a and 3b are plots of the mean and the standard deviation of the value function estimated using 1000 samples for males and females. The plots indicate that varying the choice of uncertainty budget, Γ, varies the tradeoff between the mean and the standard deviation, with the IM for transition probabilities having lowest standard deviation and

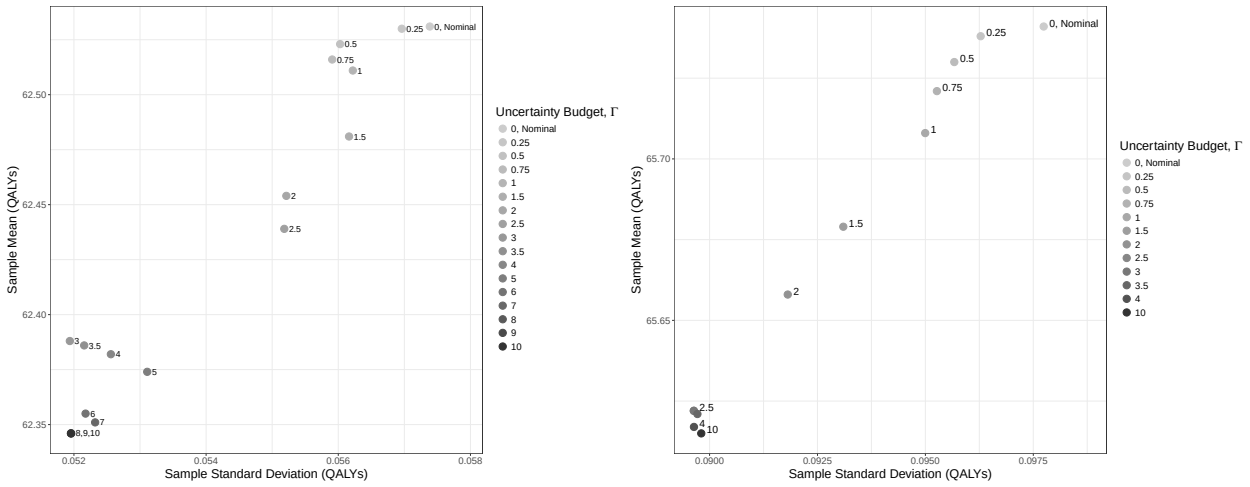


(a) Males



(b) Females

Figure 2 **Violin plots illustrating the distribution of sampled expected QALYs and theoretical worst-case expected QALYs for the MDP-optimal policy (i.e. the RMDP-TM-optimal policy with $\Gamma = 0$), and RMDP-TM-optimal policies for various choices of $\Gamma > 0$, and for varying confidence levels, $1 - \alpha$.**



(a) Males (b) Females

Figure 3 Plots illustrating the mean and the standard deviation of the value function estimated using 1000 samples for males and females for various choices of uncertainty budget, Γ, with $\alpha = 0.01$

the lowest sample mean and the MDP-optimal policy (i.e., $\Gamma = 10$) having highest sample mean but also the highest standard deviation. Thus, varying Γ provide a means to control the tradeoff between mean and standard deviation of the value function estimates.

4.3. Treatment Policy Structure Comparison

In the previous section we presented results for the value functions. In this section we present results illustrating difference in the policies themselves. Figures 4a and 4b show the treatment actions for all 10 HbA1c states for 15 years following diagnosis of type 2 diabetes for males and females, respectively. Since the prior treatment decision is part of the state space, there are 4 figures for each policy, corresponding to the most recent treatment histories (no prior treatment, treated with Metformin, treated with Sulfonylurea, and treated with Metformin and Sulfonylurea). Since insulin is the final treatment option in the model, and once on insulin, patients will continue on insulin for the remaining time, initiation of insulin completes the decision process. The figures on the left are for the nominal MDP and the figures on the right are for the RMDP-TM with $\Gamma = 10$, i.e., the interval model verison of RMDP-TM. Figures show a general trend of treating more aggressively, i.e., adding the second line medication Sulfonylurea or initiating insulin sooner. Previous work (Kaufman et al. 2011) had similar observations that the therapy is initiated sooner for the robust living-donor liver transplantation model. For the nominal MDP optimal policies, the use of metformin (an oral medication) is quite prominent. For

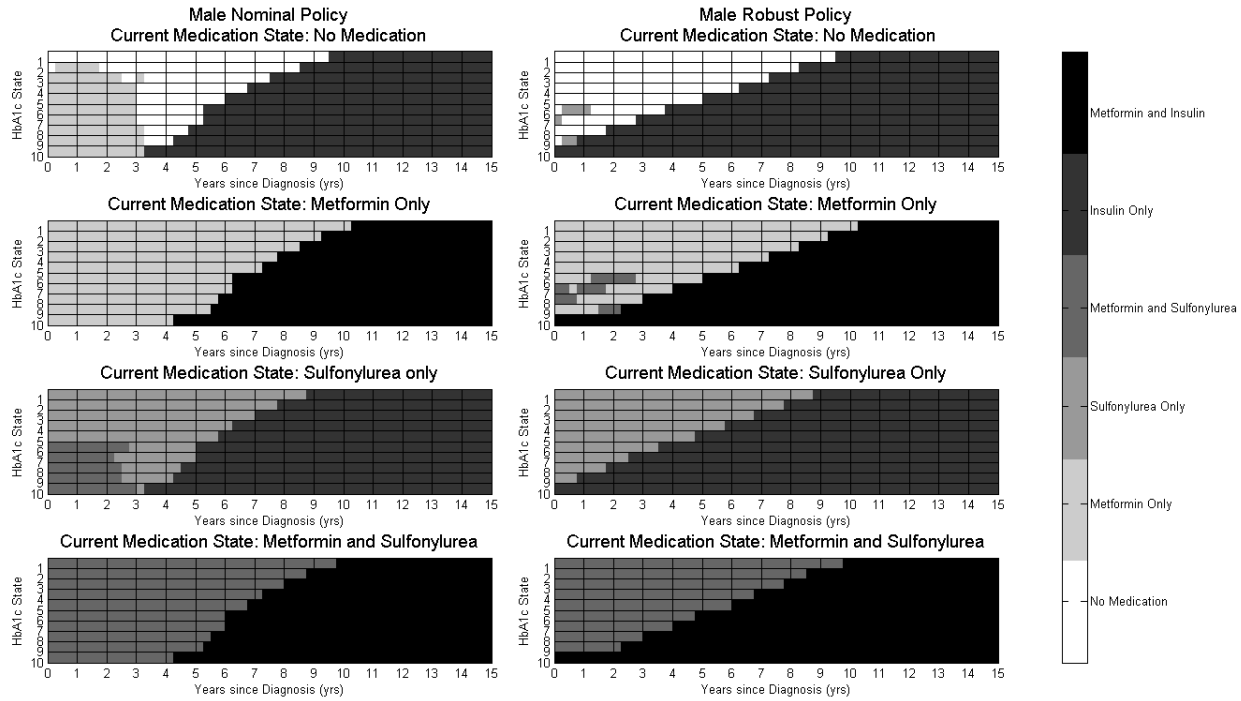
the RMDP-TM optimal policies, progression to insulin occurs much more quickly with the goal of avoiding significant complications of type 2 diabetes. It is worth noting that both of these strategies have been suggested and are debated in the endocrinology community. Although not provided for brevity, varying the choice of Γ provides a means to adjust the aggressiveness of treatment that mirrors the variation seen in the value function estimates of the previous section.

4.4. Analysis of the time-invariant counterpart of the RMDP-TM for the glycemic control problem

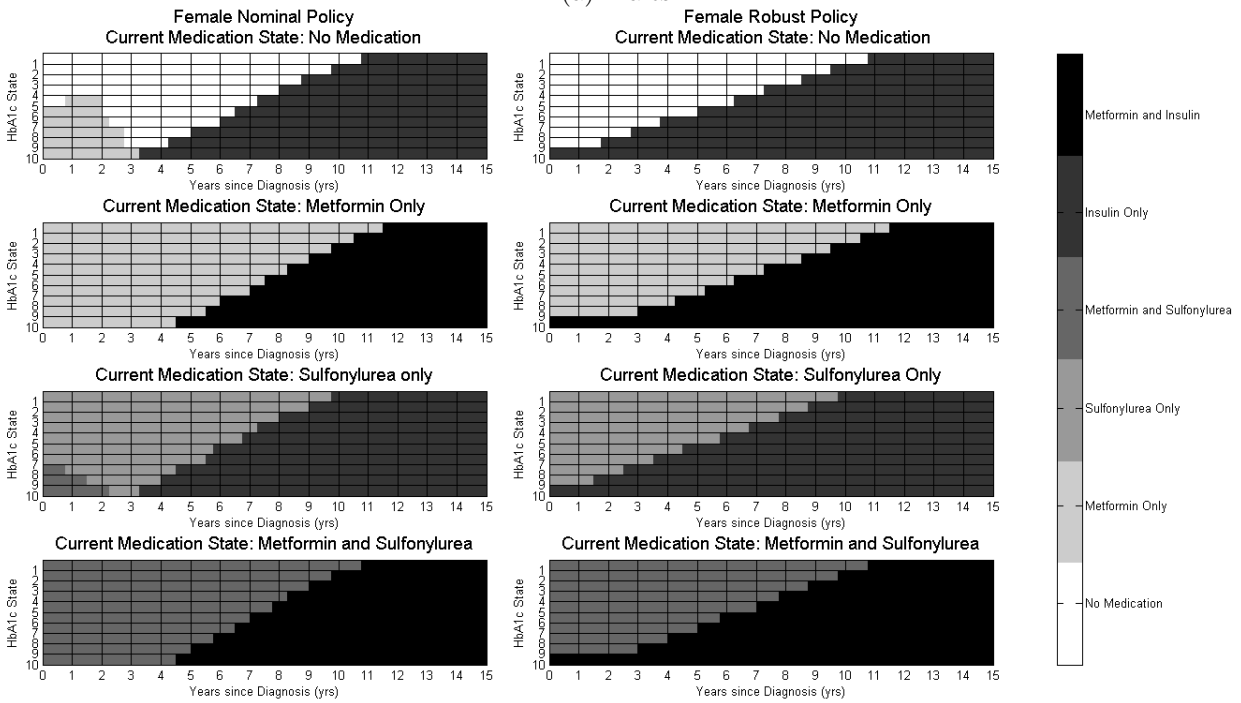
We found that the MLEs of the TPMs for both genders did not satisfy the IFR assumption exactly. However, based on the following worst-case violation measurement (Alagoz et al. 2004):

$$\epsilon = \max_{t\in\mathcal{T}} \max_{j\in\{1,\cdots,|\mathcal{L}|-1\}} \max_{i\in\{1,\cdots,|\mathcal{L}|\}} \sum_{s=i}^{|\mathcal{L}|} \left[q_{t,\ell_t(j)}(\ell_{t+1}(s)) - q_{t,\ell_t(j+1)}(\ell_{t+1}(s)) \right], \tag{21}$$

the worst-case violation was relatively small (0.125 for females and 0.1191 for males). We also calculated the NIWC TPM for males and females, and found that NIWC TPMs for both genders did not satisfy the IFR assumption exactly either. The violation is 0.0898 for female patients and 0.1129 for male patients; therefore, condition (V) in Proposition 3 did not hold. We observed that when $\Gamma = |\mathcal{L}|$, the optimal policy of nature was not stationary; therefore, this glycemic control instance of the RMDP-TM did provide the optimal solution to its time-invariant counterpart. However, we found that the optimal policy of nature was stationary for two-thirds of the 270 feasible tuples of living states, medication states, and medication initiation actions. For the remaining third of the tuples, nature's policy was stationary for 78.5%-81.7% of the decision epochs for men and 74.8%-78.8% of the decision epochs for women. Although the policy of nature was not completely stationary, we were able to generate an upper bound on the value function of the time-invariant RMDP-TM by fixing nature's policy to be $(Q_{T-1}^{\mathrm{NIWC},|\mathcal{L}|}, Q_{T-1}^{\mathrm{NIWC},|\mathcal{L}|}, \ldots, Q_{T-1}^{\mathrm{NIWC},|\mathcal{L}|})$ where $Q_{T-1}^{\mathrm{NIWC},|\mathcal{L}|}$ is nature's optimal policy for the last decision epoch in the time-variant RMDP-TM. For men, we found that this upper bound was 62.30 QALYs and for women the upper bound was 65.69 QALYs, which are tighter upper bounds than simply using the results of the respective nominal problems for men and women. Thus, even though the conditions in Proposition 3 were not strictly satisfied, solving a time-invariant problem using $Q_{T-1}^{\mathrm{NIWC},|\mathcal{L}|}$ as nature's policy in every decision epoch can provide some insights into the time-invariant counterpart of the glycemic control problem.



(a) Males



(b) Females

Figure 4 **The MDP-optimal policy (i.e. the RMDP-optimal policy with $\Gamma = 0$), and RMDP-TM optimal policies with $\Gamma = 10$.**

Although we were not able to use Proposition 3 to solve the time-invariant problem directly the property would hold for other applications that strictly meet the IFR condition, and thus Proposition 3 identifies a special class of treatment problems for which the computationally intractable time-invariant form of the RMDP can be solved easily.

5. Conclusions

We presented an RMDP model that can fit a broad range of medical treatment decisions in which there is uncertainty in transition probabilities. The interval model version of RMDP-TM model can be solved efficiently. Moreover, sufficient conditions exist under which the optimal solution to the (easy) time varying model automatically solves the much harder time invariant model. In practice the general version of latter problem is NP-hard (Wiesemann et al. 2013) but the special case we identified can be solved in polynomial time.

The IMUB version of RMDP-TM makes a novel connection between stochastic dynamic programming and the robust optimization literature by incorporating an uncertainty budget-based formulation in the MDP setting. We proved that the RMDP-TM remains computationally tractable with including the uncertainty budget constraint and the numerical experiments showed that it can control the tradeoff between the nominal performance and the worst-case performance. Moreover, changes to the budget can control other performance measures such as mean and variance of model outcomes.

We demonstrated the feasibility of solving the RMDP-TM for a practical instance of an important medical decision making problem in the context of type 2 diabetes. We illustrated how varying choices of confidence level and the uncertainty budget impacts health performance measures for norminal and RMDP-TM optimal policies. Based on our simulation study, we found that the theoretical worst-case performance may be much worse than what is observed using Monte-Carlo sampling suggesting that in some applications focusing on the theoretical worst-case may be overly conservative.

The RMDP-TM with IMUB proposed in this paper is based on the rectangular assumption on TPMs that may not fit all contexts, however, it is so far the only condition which makes RMDP solvable for large practical model instances. Future methodological research in the RMDP domain could focus on generating approximations and solution methods for solving RMDP without this assumption to better understanding the impact of this assumption.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Number CMMI-1462060 (Denton) and Grant Number DGE-1256260 (Steimle). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors would like to thank Nilay Shah and Steven Smith from Mayo Clinic for valuable feedback from clinical perspective. The authors also thank Murat Kurt for helpful comments on an early version of this manuscript.

References

Alagoz, Oguzhan, Lisa M. Maillart, Andrew J. Schaefer, Mark S. Roberts. 2004. The optimal timing of living-donor liver transplantation. *Management Science* **50** 1420–1430.

Alagoz, Oguzhan, Lisa M. Maillart, Andrew J. Schaefer, Mark S. Roberts. 2007. Choosing among living-donor and cadaveric livers. *Management Science* **53** 1702–1715.

American Diabetes Association. 2017. Standards of medical care in diabetes 2017. *Diabetes Care* **40** S1–S64.

Arias, Elizabeth. 2011. United States Life Tables, 2007. *National Vital Statistics Reports* **59**.

Barlow, Richard E., Frank Proschan. 1965. *Mathmatical theory of reliability*. Wiley.

Bertsimas, Dimitris, David B. Brown, Constantine Caramanis. 2011. Theory and applications of robust optimization. *SIAM Review* **53** 464–501.

Bertsimas, Dimitris, Melvyn Sim. 2004. The price of robustness. *Operations Research* **52** 35–53.

Centers for Disease Control and Prevention. 2012. Deaths, percent of total deaths, and death rates for the 15 leading causes of death in 10-year age groups, by race and sex: United states, 1999-2007.

Centers for Disease Control and Prevention. 2013. Age at diagnosis of diabetes among adult incident cases aged 18-79 years.

Clarke, P., A. Gray, A. Briggs, A. Farmer, P. Fenn, R. Stevens, D. Matthews, I. Stratton, R. Holman. 2004. A model to estimate the lifetime health outcomes of patients with type 2 diabetes: the united kingdom prospective diabetes study (UKPDS) outcomes model (UKPDS no. 68). *Diabetologia* **47** 1747–1759.

Craig, Bruce A., Peter P.G Sendi. 2002. Estimation of the transition matrix of a discrete-time Markov chain. *Health Economics* **11** 33–43.

Fitzpatrick, Simon, Alastair Scott. 1987. Quick simultaneous confidence intervals for multinomial proportions. *Journal of the American Statistical Association* **82** 875–878.

Gold, Marthe R., David Stevenson, Dennis G. Fryback. 2002. HALYS and QALYS and DALYS, Oh My: Similarities and Differences in Summary Measures of Population Health. *Annual Review of Public Health* **23** 115–134.

Gold, Ruth Z. 1963. Tests auxiliary to χ^2 tests in a markov chain. *The Annals of Mathematical Statistics* **34** 56–74.

Goodman, Leo A. 1965. On simultaneous confidence intervals for multinomial proportions. *Technometrics* **7** 247–254.

Iyengar, Garud N. 2005. Robust dynamic programming. *Mathematics of Operations Research* **30** 257–280.

Kaufman, David L., Andrew J. Schaefer. 2012. Robust modified policy iteration. *INFORMS Journal on Computing* 1–15.

Kaufman, David L., Andrew J. Schaefer, Mark S. Roberts. 2011. Living-donor liver transplantation timing under ambiguous health state transition probabilities (extended abstract). *Proceedings of the 2011 Manufacturing and Service Operations Management (MSOM) Conference* .

Kurt, Murat, Brian T. Denton, Andrew J. Schaefer, Nilay D. Shah, Steven A. Smith. 2011. The structure of optimal statin initiation policies for patients with type 2 diabetes. *IIE Transactions on Healthcare Systems Engineering* **1** 49–65.

Mannor, Shie, Ofir Mebel, Huan Xu. 2016. Robust mdps with k-rectangular uncertainty. *Mathematics of Operations Research* **41** 1484–1509.

Mason, Jennifer E., Brian T. Denton, Nilay D. Shah, S. A. Smith. 2014. Optimizing the simultaneous management of blood pressure and cholesterol for type 2 diabetes patients. *European Journal of Operational Research* 727–738.

Nilim, Arnab, Laurent El Ghaoui. 2005. Robust control of markov decision processes with uncertain transition matrices. *Operations Research* **53** 780–798.

Pignone, Michael, Stephanie Earnshaw, Jeffrey A. Tice, Mark J. Pletcher. 2006. Aspirin, statins, or both drugs for the primary prevention of coronary heart disease events in men: A costutility analysis. *Annals of Internal Medicine* **144** 326–336.

Puterman, Martin L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., Hoboken, New Jersey, USA.

Satia, Jay K., Roy E. Lave Jr. 1973. Markovian decision processes with uncertain transition probabilities. *Operations Research* **21** 728–740.

Shechter, Steven M., Matthew D. Bailey, Andrew J. Schaefer, Mark S. Roberts. 2008. The optimal time to initiate HIV therapy under ordered health states. *Operations Research* **56** 20–33.

Sinha, Anushua, Mangala Rajan, Thomas Hoerger, Len Pogach. 2010. Costs and consequences associated with newer medications for glycemic control in type 2 diabetes. *Diabetes Care* **33** 695–700.

Sison, Cristina P., Joseph Glaz. 1995. Simultaneous confidence intervals and sample size determination for multinomial proportions. *Journal of the American Statistical Association* **90** 366–369.

UKPDS Group. 1995. U.K. prospective diabetes study 16. Overview of 6 years' therapy of type II diabetes: a progressive disease. U.K. Prospective Diabetes Study Group. *Diabetes* **44** 1249–1258.

White III, Chelsea C., Hany K. Eldeib. 1994. Markov decision processes with imprecise transition probabilities. *Operations Research* **42** 739–749.

Wiesemann, Wolfram, Daniel Kuhn, Berc Rustem. 2013. Robust markov decision processes. *Mathematics of operations research* **38** 153–183.

Zhang, Yuanhui, Rozalina G. McCoy, Jennifer E. Mason, Steve A. Smith, Nilay D. Shah, Brian T. Denton. 2014. Second-line agents for glycemic control for type 2 diabetes: Are newer agents better? *Diabetes Care* .

Zhang, Yuanhui, Haipeng Wu, Brian T. Denton, James R. Wilson, Jennifer M. Lobo. 2016. Probabilistic sensitivity analysis on markov models with uncertain transition probabilities: An application in evaluating treatment decisions for type 2 diabetes. `http://btdenton.engin.umich.edu/wp-content/uploads/sites/138/2016/12/Zhang-2016.pdf`.

Appendix

A. Proof of Propositions

Proof of Proposition 1 Since it is clear that $\mathcal{Q}^{\text{IMUB-NLP}} \subseteq \mathcal{Q}^{\text{IMUB-LP}}$, to show that $\mathcal{Q}^{\text{IMUB-NLP}} \equiv \mathcal{Q}^{\text{IMUB-LP}}$, it suffices to show that $\mathcal{Q}^{\text{IMUB-NLP}} \supseteq \mathcal{Q}^{\text{IMUB-LP}}$. To show this, consider $\forall q_{t,\ell_t}(\ell_{t+1}) = \hat{q}_{t,\ell_t}(\ell_{t+1}) + z^l_{t,\ell_t}(\ell_{t+1})\delta^l_{t,\ell_t}(\ell_{t+1}) + z^u_{t,\ell_t}(\ell_{t+1})\delta^u_{t,\ell_t}(\ell_{t+1}) \in \mathcal{Q}^{\text{IMUB-LP}}$. We define the quantities $\tilde{z}^l_{t,\ell_t}(\ell_{t+1}), \tilde{z}^u_{t,\ell_t}(\ell_{t+1})$ as follows:

$$\begin{aligned}
\tilde{z}^l_{t,\ell_t}(\ell_{t+1}) &= \frac{1}{\delta^l_{t,\ell_t}(\ell_{t+1})}\left(\delta^l_{t,\ell_t}(\ell_{t+1})z^l_{t,\ell_t}(\ell_{t+1}) - \delta^u_{t,\ell_t}(\ell_{t+1})z^u_{t,\ell_t}(\ell_{t+1})\right)^+ \\
\tilde{z}^u_{t,\ell_t}(\ell_{t+1}) &= \frac{1}{\delta^u_{t,\ell_t}(\ell_{t+1})}\left(\delta^u_{t,\ell_t}(\ell_{t+1})z^u_{t,\ell_t}(\ell_{t+1}) - \delta^l_{t,\ell_t}(\ell_{t+1})z^l_{t,\ell_t}(\ell_{t+1})\right)^+
\end{aligned}$$

Based on the definition above, we know that $\tilde{z}^l_{t,\ell_t}(\ell_{t+1}) \geq 0$ and $\tilde{z}^u_{t,\ell_t}(\ell_{t+1}) \geq 0$. The reminder is to show that $(q_{t,\ell_t}(\ell_{t+1}), \tilde{z}^l_{t,\ell_t}(\ell_{t+1}), \tilde{z}^u_{t,\ell_t}(\ell_{t+1}))$ satisfies the six constraints defining $\mathcal{Q}^{\text{IMUB-NLP}}$. For the first constraint:

$$\begin{aligned}
&\hat{q}_{t,\ell_t}(\ell_{t+1}) - \delta^l_{t,\ell_t}(\ell_{t+1})\tilde{z}^l_{t,\ell_t}(\ell_{t+1}) + \delta^u_{t,\ell_t}(\ell_{t+1})\tilde{z}^u_{t,\ell_t}(\ell_{t+1}) \\
=\ &\hat{q}_{t,\ell_t}(\ell_{t+1}) - \left(\delta^l_{t,\ell_t}(\ell_{t+1})z^l_{t,\ell_t}(\ell_{t+1}) - \delta^u_{t,\ell_t}(\ell_{t+1})z^u_{t,\ell_t}(\ell_{t+1})\right)^+ \\
&+ \left(-\delta^l_{t,\ell_t}(\ell_{t+1})z^l_{t,\ell_t}(\ell_{t+1}) + \delta^u_{t,\ell_t}(\ell_{t+1})z^u_{t,\ell_t}(\ell_{t+1})\right)^+ \\
=\ &\hat{q}_{t,\ell_t}(\ell_{t+1}) - \delta^l_{t,\ell_t}(\ell_{t+1})z^l_{t,\ell_t}(\ell_{t+1}) + \delta^u_{t,\ell_t}(\ell_{t+1})z^u_{t,\ell_t}(\ell_{t+1}) \\
=\ &q_{t,\ell_t}(\ell_{t+1})
\end{aligned}$$

The second and the sixth constraints only depends on $q_{t,\ell_t}(\ell_{t+1})$. For the third and the fifth constraints, we know that

$$\begin{aligned}
\tilde{z}^l_{t,\ell_t}(\ell_{t+1}) &= \frac{1}{\delta^l_{t,\ell_t}(\ell_{t+1})}\left(\delta^l_{t,\ell_t}(\ell_{t+1})z^l_{t,\ell_t}(\ell_{t+1}) - \delta^u_{t,\ell_t}(\ell_{t+1})z^u_{t,\ell_t}(\ell_{t+1})\right)^+ \\
&\leq \frac{1}{\delta^l_{t,\ell_t}(\ell_{t+1})}\left(\delta^l_{t,\ell_t}(\ell_{t+1})z^l_{t,\ell_t}(\ell_{t+1}) - 0\right)^+ \\
&= z^l_{t,\ell_t}(\ell_{t+1}). \\
\tilde{z}^u_{t,\ell_t}(\ell_{t+1}) &= \frac{1}{\delta^u_{t,\ell_t}(\ell_{t+1})}\left(\delta^u_{t,\ell_t}(\ell_{t+1})z^u_{t,\ell_t}(\ell_{t+1}) - \delta^l_{t,\ell_t}(\ell_{t+1})z^l_{t,\ell_t}(\ell_{t+1})\right)^+ \\
&\leq \frac{1}{\delta^u_{t,\ell_t}(\ell_{t+1})}\left(\delta^u_{t,\ell_t}(\ell_{t+1})z^u_{t,\ell_t}(\ell_{t+1}) - 0\right)^+ \\
&= z^u_{t,\ell_t}(\ell_{t+1}).
\end{aligned}$$

Therefore, $0 \leq \tilde{z}^l_{t,\ell_t}(\ell_{t+1}) \leq 1$, $0 \leq \tilde{z}^u_{t,\ell_t}(\ell_{t+1}) \leq 1$, and $\tilde{z}^u_{t,\ell_t}(\ell_{t+1}) + \tilde{z}^u_{t,\ell_t}(\ell_{t+1}) \leq z^l_{t,\ell_t}(\ell_{t+1}) + z^u_{t,\ell_t}(\ell_{t+1}) \leq \Gamma_{t,\ell_t}$. For the fourth constraint, i.e., the nonlinear constraint,

$$\begin{aligned}
&\tilde{z}^l_{t,\ell_t}(\ell_{t+1}) \cdot \tilde{z}^u_{t,\ell_t}(\ell_{t+1}) \\
=\ &\frac{1}{\delta^l_{t,\ell_t}(\ell_{t+1})}\left(\delta^l_{t,\ell_t}(\ell_{t+1})z^l_{t,\ell_t}(\ell_{t+1}) - \delta^u_{t,\ell_t}(\ell_{t+1})z^u_{t,\ell_t}(\ell_{t+1})\right)^+ \\
&\cdot\frac{1}{\delta^u_{t,\ell_t}(\ell_{t+1})}\left(\delta^u_{t,\ell_t}(\ell_{t+1})z^u_{t,\ell_t}(\ell_{t+1}) - \delta^l_{t,\ell_t}(\ell_{t+1})z^l_{t,\ell_t}(\ell_{t+1})\right)^+ \\
=\ &0.
\end{aligned}$$

Therefor, $\mathcal{Q}^{\text{IMUB-NLP}} \equiv \mathcal{Q}^{\text{IMUB-LP}}$ and the proof is complete. □

Proof of Proposition 2: We rewrite the inner problem (6) with the IM model in the following form:

$$\text{(LP.1)}\quad \begin{aligned}
\min\quad & \sum_{i=1}^{|\mathcal{L}|} c_i(y_i + y^l_i) \\
\text{s.t.}\quad & 0 \leq y_i \leq y^u_i - y^l_i, && i = 1, \ldots, |\mathcal{L}|, \\
& \sum_{i=1}^{|\mathcal{L}|}(y_i + y^l_i) = 1,
\end{aligned}$$

where y_i, $\forall i \in \{1, \cdots, |\mathcal{L}|\}$, are the decision variables, $c_i \geq 0$, $\forall i \in \{1, \cdots, |\mathcal{L}|\}$, are the coefficients of the

objective function, y_i^l and y_i^u, $\forall i \in \{1, \cdots, |\mathcal{L}|\}$ are lower and upper bounds of the transition probabilities, respectively. We assume that $0 \le y_i^l < y_i^u$ without loss of generality. The associated dual problem is:

$$\text{(Dual of LP.1)} \quad \begin{aligned} \max \quad & \sum_{i=1}^{|\mathcal{L}|} c_i y_i^l + \sum_{i=1}^{|\mathcal{L}|} (y_i^u - y_i^l)\alpha_i + (1 - \sum_{i=1}^{|\mathcal{L}|} y_i^l)\beta \\ \text{s.t.} \quad & \alpha_i + \beta \le c_i, \forall i = 1, \ldots, |\mathcal{L}| \\ & \alpha_i \le 0, \forall i = 1, \ldots, |\mathcal{L}| \\ & \beta \in \mathbb{R} \end{aligned}$$

One readily verifies that $\alpha_i^* = \min\{c_i - \beta, 0\}$, $\forall i \in \{1, \ldots, |\mathcal{L}|\}$. We thus have to solve the unrestricted, one-dimensional problem:

$$\max. \quad (1 - \sum_{i=1}^{|\mathcal{L}|} y_i^l)\beta + \sum_{i=1}^{|\mathcal{L}|} (y_i^u - y_i^l) \min\{c_i - \beta, 0\} + \sum_{i=1}^{|\mathcal{L}|} c_i y_i^l$$

The optimal solution must be attained at a 'kink' where $\beta = c_i$ for some i, so we just have to evaluate the objective function for each choice of $\beta = c_i$, and select the best one. The complexity of the Algorithm 1 depends on the for loop from line 4–11, therefore, it has complexity $O(|\mathcal{L}|)$. $\square$

The following Lemma 1 will be used to prove the Proposition 3. Lemma 1 is similar to the Lemma 4.7.2 in Puterman (1994), however, Lemma 4.7.2 in Puterman (1994) is for the infinite-horizon case, and for nondecreasing sequence $\{v_j\}_{j=0,1,\ldots}$.

LEMMA 1. *Let $\{x_j\}$, $\{x'_j\}$ be real-value non-negative sequences satisfying*

$$\sum_{j=n}^{N} x_j \ge \sum_{j=n}^{N} x'_j$$

for all n, with equality holding for $n = 0$. Suppose $v_{j+1} \le v_j$ for $j = 0, 1, \ldots, N$, then

$$\sum_{j=0}^{N} v_j x_j \le \sum_{j=0}^{N} v_j x'_j$$

Proof of Lemma 1 Let n be arbitrary and $v_{-1} = 0$. Then

$$\begin{aligned} \sum_{j=0}^{N} v_j x_j &= \sum_{j=0}^{N} x_j \sum_{i=0}^{j} (v_i - v_{i-1}) = \sum_{j=0}^{N} (v_j - v_{j-1}) \sum_{i=j}^{N} x_i = \sum_{j=1}^{N} (v_j - v_{j-1}) \sum_{i=j}^{N} x_i + v_0 \sum_{i=0}^{N} x_i \\ &\le \sum_{j=1}^{N} (v_j - v_{j-1}) \sum_{i=j}^{N} x'_i + v_0 \sum_{i=0}^{N} x_i \quad (\because v_j - v_{j-1} \le 0, \forall j = 1, \ldots, N) \\ &= \sum_{j=1}^{N} (v_j - v_{j-1}) \sum_{i=j}^{N} x'_i + v_0 \sum_{i=0}^{N} x'_i \quad (\because \sum_{j=0}^{N} x_j = \sum_{j=0}^{N} x'_j) \\ &= \sum_{j=0}^{N} (v_j - v_{j-1}) \sum_{i=j}^{N} x'_i = \sum_{j=0}^{N} x'_j \sum_{i=0}^{j} (v_i - v_{i-1}) = \sum_{j=0}^{N} v_j x'_j \end{aligned}$$

Therefore, $\sum_{j=0}^{N} v_j x_j \le \sum_{j=0}^{N} v_j x'_j$ holds.

Proof of Proposition 3: First, we show that given Conditions (I) and (II), the NIWC TPMs, $\left\{Q_t^{\text{NIWC},|\mathcal{L}|}\right\}_{t\in\mathcal{T}\setminus\{T\}}$, defined by Definition 1, are the same for all time epochs. By Condition (I) and Proposition 2, each row of the NIWC TPM, $Q_t^{\text{NIWC},|\mathcal{L}|}$ can be generated by using Algorithm 1. The order of the coefficients in the objective function of the problem (20) is fixed (i.e. nonincreasing in health states), therefore, as shown in Algorithm 1, the optimal solution, $\boldsymbol{q}_{t,\ell}^{\text{NIWC},|\mathcal{L}|}, \forall\, \ell\in\mathcal{L}$, is only determined by the feasible region of the problem (20) (i.e, the uncertainty set, $\mathcal{Q}_{t,\ell}^{\text{IM}}$). By Condition (II) we know $\mathcal{Q}_{t,\ell}^{\text{IM}} = \mathcal{Q}_{t',\ell}^{\text{IM}}, \forall \ell\in\mathcal{L}, t, t'\in\mathcal{T}\setminus\{T\}$. Therefore,

$$\boldsymbol{q}_{t,\ell}^{\text{NIWC},|\mathcal{L}|} = \boldsymbol{q}_{t',\ell}^{\text{NIWC},|\mathcal{L}|}, \quad \forall t, t'\in\mathcal{T}\setminus\{T\}, \ell\in\mathcal{L}.$$

It follows that

$$Q_1^{\text{NIWC},|\mathcal{L}|} = Q_2^{\text{NIWC},|\mathcal{L}|} = \cdots = Q_{T-1}^{\text{NIWC},|\mathcal{L}|} \tag{22}$$

Next, we prove results (a) and (b) hold by induction. For the base case $t=T$, $v_T^{\text{RMDP-TM}}(\ell_T, \boldsymbol{m}_T) = r_t(\ell_T, \boldsymbol{m}_T)$ is nonincreasing in ℓ_T, $\forall\, \boldsymbol{m}_T\in\mathcal{M}$ by Condition (III). Now assume that $v_k^{\text{RMDP-TM}}(\ell_k, \boldsymbol{m}_k)$ is nonincreasing in ℓ_k, $\forall\, \boldsymbol{m}_k\in\mathcal{M}$ and $k\in\{t+1, t+2, \ldots, T-1\}$, then we need to prove $v_t^{\text{RMDP-TM}}(\ell_t, \boldsymbol{m}_t)$ is nonincreasing in ℓ_t, $\forall\, \boldsymbol{m}_t\in\mathcal{M}$. Given the base case and induction hypothesis, the optimal decision of nature for $\tau\in\{t, t+1, \ldots, T-1\}$ is the NIWC TPM, $Q_\tau^{\text{NIWC},|\mathcal{L}|}$, of the uncertainty set $\mathcal{Q}_\tau$ based on Definition 1. Combined Condition (V) and Equation (22) we know that $Q_t^{\text{NIWC}}(|\mathcal{L}|)$ has the IFR property. By Lemma 1, we have

$$\sum_{\ell_{t+1}\in\mathcal{L}} q_{t,\ell_t}^{\text{NIWC}}(\ell_{t+1}) v_{t+1}^{\text{RMDP-TM}}(\ell_{t+1}, \boldsymbol{m}_{t+1}) \geq \sum_{\ell_{t+1}\in\mathcal{L}} q_{t,\ell_t'}^{\text{NIWC}}(\ell_{t+1}) v_{t+1}^{\text{RMDP-TM}}(\ell_{t+1}, \boldsymbol{m}_{t+1}), \tag{23}$$

if the health state $\ell_t\in\mathcal{L}$ is order such that a patient in health state ℓ_t is healthier than a patient in health state ℓ_t'. Combined with the conditions that the one-period reward at decision epoch t is nonincreasing in ℓ_t for any $\boldsymbol{m}_t\in\mathcal{M}$ (Condition (III)), and the probability of entering absorbing state is nondecreasing in ℓ_t for any $\boldsymbol{m}_t\in\mathcal{M}$, $\boldsymbol{\alpha}_t\in\mathcal{A}_t(\ell_t, \boldsymbol{m}_t)$ (Condition (IV)), we have $v_t^{\text{RMDP-TM}}(\ell_t, \boldsymbol{m}_t)$ is nonincreasing in ℓ_t for any $\boldsymbol{m}_t\in\mathcal{M}$. Therefore, the induction hypothesis is satisfied, and the result (a) follows.

Based on result (a) that the optimal value function is nonincreasing with respect to health state for any decision epochs, the optimal decision of nature for any decision epoch $t\in\mathcal{T}\setminus\{T\}$ is the NIWC TPM, $Q_t^{\text{NIWC},|\mathcal{L}|}$, of the uncertainty set $\mathcal{Q}_t$ based on Definition 1. Combined with Equation (22), we have the optimal policy of nature is $(Q_{T-1}^{\text{NIWC},|\mathcal{L}|}, \ldots, Q_{T-1}^{\text{NIWC},|\mathcal{L}|})$ which is stationary. □

B. Parameters Values for the Case Study

Table 1 Glycosylated hemoglobin (HbA1c) used in the RMDP-TM for women. HbA1c range definition at diagnosis, the mean natural HbA1c values for each HbA1c state at diagnosis (prior to initiating medication), the initial HbA1c distributions at diagnosis, and 3-month HbA1c transition probability matrices for women.

		HbA1c State									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
HbA1c Range		<6	[6,6.5)	[6.5,7)	[7,7.5)	[7.5,8)	[8,8.5)	[8.5,9)	[9,9.5)	[9.5,10)	10
Mean HbA1c value (%)		5.7	6.25	6.74	7.24	7.73	8.23	8.73	9.22	9.72	11.73
Initial HbA1c Distribution		0.0771	0.1543	0.2125	0.18	0.1105	0.0848	0.0502	0.035	0.0273	0.0683
TPM	**HbA1c state 1**	0.6471	0.3529	0	0	0	0	0	0	0	0
	HbA1c state 2	0.1800	0.5200	0.2200	0.0600	0.0200	0	0	0	0	0
	HbA1c state 3	0.0435	0.1957	0.4783	0.2174	0.0652	0	0	0	0	0
	HbA1c state 4	0.0192	0.0577	0.2500	0.3846	0.1923	0.0769	0.0192	0	0	0
	HbA1c state 5	0.0323	0	0.1935	0.2903	0.2258	0.1935	0.0323	0	0.0323	0
	HbA1c state 6	0	0	0.0370	0.1852	0.1852	0.2963	0.2222	0.0370	0.0370	0
	HbA1c state 7	0	0.0588	0	0.0588	0.2353	0.2353	0.1765	0.1176	0.1176	0
	HbA1c state 8	0	0	0	0	0	0.2500	0.2500	0	0.5000	0
	HbA1c state 9	0	0	0	0	0	0.1250	0.1250	0.2500	0.3750	0.1250
	HbA1c state 10	0	0	0	0.0500	0.1500	0.0500	0.1000	0	0.2000	0.4500

Table 2 Glycosylated hemoglobin (HbA1c) used in the RMDP-TM for men. HbA1c range definition at diagnosis, the mean natural HbA1c values for each HbA1c state at diagnosis (prior to initiating medication), the initial HbA1c distributions at diagnosis, and 3-month HbA1c transition probability matrices for men.

		HbA1c State									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
HbA1c Range		<6	[6,6.5)	[6.5,7)	[7,7.5)	[7.5,8)	[8,8.5)	[8.5,9)	[9,9.5)	[9.5,10)	10
Mean HbA1c value (%)		5.69	6.25	6.73	7.24	7.74	8.24	8.74	9.21	9.73	11.59
Initial HbA1c Distribution		0.0694	0.1388	0.1968	0.1626	0.1138	0.0919	0.0619	0.0424	0.0328	0.0896
TPM	**HbA1c state 1**	0.6667	0.2444	0.0889	0	0	0	0	0	0	0
	HbA1c state 2	0.1346	0.5000	0.3654	0	0	0	0	0	0	0
	HbA1c state 3	0.0794	0.2222	0.3810	0.2857	0	0.0317	0	0	0	0
	HbA1c state 4	0.0175	0.0877	0.2807	0.3860	0.1579	0.0702	0	0	0	0
	HbA1c state 5	0	0.0465	0.2093	0.2326	0.2326	0.1628	0.0930	0.0233	0	0
	HbA1c state 6	0	0	0.0800	0.0800	0.2000	0.3600	0.2000	0.0400	0	0.0400
	HbA1c state 7	0.1071	0	0.0357	0.1071	0.0714	0.1429	0.2143	0.2500	0.0357	0.0357
	HbA1c state 8	0	0.0833	0	0.0833	0.2500	0.1667	0	0.2500	0.0833	0.0833
	HbA1c state 9	0.0556	0.0556	0	0.0556	0.1667	0.1111	0.1111	0.2222	0.0556	0.1667
	HbA1c state 10	0	0	0.0588	0.1176	0.0588	0.1765	0.1176	0.0588	0.0882	0.3236

Table 3 Left-hand-side maximum deviation of the TPM in RMDP-TM for women.

	HbA1c State									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
HbA1c State 1	0.2353	0.2353	0	0	0	0	0	0	0	0
HbA1c State 2	0.1600	0.1600	0.1600	0.0600	0.0200	0	0	0	0	0
HbA1c State 3	0.0435	0.1739	0.1739	0.1739	0.0652	0	0	0	0	0
HbA1c State 4	0.0192	0.0577	0.1538	0.1538	0.1538	0.0769	0.0192	0	0	0
HbA1c State 5	0.0323	0	0.1935	0.1935	0.1935	0.1935	0.0323	0	0.0323	0
HbA1c State 6	0	0	0.0370	0.1852	0.1852	0.1852	0.1852	0.0370	0.0370	0
HbA1c State 7	0	0.0588	0	0.0588	0.2353	0.2353	0.1765	0.1176	0.1176	0
HbA1c State 8	0	0	0	0	0	0.2500	0.2500	0	0.2500	0
HbA1c State 9	0	0	0	0	0	0.1250	0.1250	0.2500	0.2500	0.1250
HbA1c State 10	0	0	0	0.0500	0.1500	0.0500	0.1000	0	0.2000	0.2000

Table 4 Right-hand-side maximum deviation of the TPM in RMDP-TM for women.

	HbA1c State									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
HbA1c State 1	0.2841	0.2841	0.2841	0.2841	0.2841	0.2841	0.2841	0.2841	0.2841	0.2841
HbA1c State 2	0.1765	0.1765	0.1765	0.1765	0.1765	0.1765	0.1765	0.1765	0.1765	0.1765
HbA1c State 3	0.1778	0.1778	0.1778	0.1778	0.1778	0.1778	0.1778	0.1778	0.1778	0.1778
HbA1c State 4	0.1837	0.1837	0.1837	0.1837	0.1837	0.1837	0.1837	0.1837	0.1837	0.1837
HbA1c State 5	0.2265	0.2265	0.2265	0.2265	0.2265	0.2265	0.2265	0.2265	0.2265	0.2265
HbA1c State 6	0.2541	0.2541	0.2541	0.2541	0.2541	0.2541	0.2541	0.2541	0.2541	0.2541
HbA1c State 7	0.2924	0.2924	0.2924	0.2924	0.2924	0.2924	0.2924	0.2924	0.2924	0.2924
HbA1c State 8	0.7153	0.7153	0.7153	0.7153	0.7153	0.7153	0.7153	0.7153	0.5000	0.7153
HbA1c State 9	0.4840	0.4840	0.4840	0.4840	0.4840	0.4840	0.4840	0.4840	0.4840	0.4840
HbA1c State 10	0.2981	0.2981	0.2981	0.2981	0.2981	0.2981	0.2981	0.2981	0.2981	0.2981

Table 5 Left-hand-side maximum deviation of the TPM in RMDP-TM for men.

	HbA1c State									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
HbA1c State 1	0.1556	0.1556	0.0889	0	0	0	0	0	0	0
HbA1c State 2	0.1346	0.1538	0.1538	0	0	0	0	0	0	0
HbA1c State 3	0.0794	0.1429	0.1429	0.1429	0	0.0317	0	0	0	0
HbA1c State 4	0.0175	0.0877	0.1579	0.1579	0.1579	0.0702	0	0	0	0
HbA1c State 5	0	0.0465	0.1628	0.1628	0.1628	0.1628	0.0930	0.0233	0	0
HbA1c State 6	0	0	0.0800	0.0800	0.2000	0.2000	0.2000	0.0400	0	0.0400
HbA1c State 7	0.1071	0	0.0357	0.1071	0.0714	0.1429	0.1786	0.1786	0.0357	0.0357
HbA1c State 8	0	0.0833	0	0.0833	0.2500	0.1667	0	0.2500	0.0833	0.0833
HbA1c State 9	0.0556	0.0556	0	0.0556	0.1667	0.1111	0.1111	0.2222	0.0556	0.1667
HbA1c State 10	0	0	0.05882	0.1176	0.0588	0.1765	0.1176	0.0588	0.0882	0.1765

Table 6 **Right-hand-side maximum deviation of the TPM in RMDP-TM for men.**

	HbA1c State									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
HbA1c State 1	0.1575	0.1575	0.1575	0.1575	0.1575	0.1575	0.1575	0.1575	0.1575	0.1575
HbA1c State 2	0.1892	0.1892	0.1892	0.1892	0.1892	0.1892	0.1892	0.1892	0.1892	0.1892
HbA1c State 3	0.1715	0.1715	0.1715	0.1715	0.1715	0.1715	0.1715	0.1715	0.1715	0.1715
HbA1c State 4	0.1666	0.1666	0.1666	0.1666	0.1666	0.1666	0.1666	0.1666	0.1666	0.1666
HbA1c State 5	0.1899	0.1899	0.1899	0.1899	0.1899	0.1899	0.1899	0.1899	0.1899	0.1899
HbA1c State 6	0.2605	0.2605	0.2605	0.2605	0.2605	0.2605	0.2605	0.2605	0.2605	0.2605
HbA1c State 7	0.2344	0.2344	0.2344	0.2344	0.2344	0.2344	0.2344	0.2344	0.2344	0.2344
HbA1c State 8	0.3656	0.3656	0.3656	0.3656	0.3656	0.3656	0.3656	0.3656	0.3656	0.3656
HbA1c State 9	0.2695	0.2695	0.2695	0.2695	0.2695	0.2695	0.2695	0.2695	0.2695	0.2695
HbA1c State 10	0.2100	0.2100	0.2100	0.2100	0.2100	0.2100	0.2100	0.2100	0.2100	0.2100