INFORMS TutORials in Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Optimization of Sequential Decision Making for Chronic Diseases: From Data to Decisions

Brian T. Denton

To cite this entry: Brian T. Denton. Optimization of Sequential Decision Making for Chronic Diseases: From Data to Decisions. *In* INFORMS TutORials in Operations Research. Published online: 19 Oct 2018; 316-348.
https://doi.org/10.1287/educ.2018.0184

https://pubsonline.informs.org/series/educ/

TUTORIALS IN OPERATIONS RESEARCH
 | ISBN 978-0-9906153-2-3
https://doi.org/10.1287/educ.2018.0184

Optimization of Sequential Decision Making for Chronic Diseases: From Data to Decisions

Brian T. Denton
Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, Michigan 48109
Contact: btdenton@umich.edu, https://orcid.org/0000-0002-6372-6066 (BTD)

Abstract Rapid advances in healthcare for chronic diseases such as cardiovascular disease, cancer, and diabetes have made it possible to detect diseases at early stages and tailor treatment based on individual patient risk factors including demographic factors and disease-specific biomarkers. However, a large number of relevant risk factors, combined with uncertainty in future health outcomes and the side effects of health interventions, makes clinical management of diseases challenging for physicians and patients. Data-driven operations research methods have the potential to help improve medical decision making by using observational data that are now routinely collected in many health systems. Optimization methods in particular, such as Markov decision processes and partially observable Markov decision processes, have the potential to improve the protracted sequential decision-making process that is common to many chronic diseases. This tutorial provides an introduction to some of the most commonly used methods for building and solving models to optimize sequential decision making. The context of chronic diseases is emphasized, but the methods apply broadly to sequential decision making under uncertainty. We pay special attention to the challenges associated with using observational data and the influence of model parameter uncertainty and ambiguity.

Keywords stochastic dynamic programming • Markov decision process • hidden Markov model • chronic disease • data analytics

1. Introduction

Chronic diseases are medical conditions that are managed over time, often for years or decades, under uncertainty about future health outcomes. They include diseases such as cardiovascular disease, cancer, and diabetes, which are collectively the most common causes of death in the United States and many developed countries (Centers for Disease Control and Prevention [28]). Rapid advances in *health interventions* such as diagnostic tests, medications, procedures, and surgery have made it possible to detect chronic diseases at early stages and target therapeutic interventions based on personal health history and risk factors. However, choosing the right intervention at a particular time requires an understanding of uncertainty in disease progression, disease outcomes, potential side effects of interventions, and future recourse decisions. Randomized trials are considered the gold standard for decision making in medicine, but they are often difficult or impossible to conduct because of their high cost, difficulty in patient recruiting, and ethical barriers. Specifically, in the case of chronic diseases, the long follow-up time needed to measure effects is a major difficulty. Furthermore, it is not practical to try to test a large number of potential policies by way of randomized trials. For these reasons, data-driven models have an important role to play in helping physicians improve treatment decisions for patients with chronic diseases.

Stochastic models are commonplace in the field of medicine where they have been used to evaluate decisions in a broad range of contexts. *Markov chains*, in particular, are among the most commonly used stochastic models for medical decision making. A keyword search of the U.S. Library of Medicine Database using PubMed from 2007 to 2017 revealed more than 7,500 articles on the topic of Markov chains. Markov chains that include decisions (i.e., Markov decision processes (MDPs)) have been applied much less; however, many applications in medicine have emerged in recent years. In the context of chronic diseases, past work employing MDPs includes liver transplant decisions (Alagoz et al. [2]), human immunodeficiency virus (HIV) treatment (Shechter et al. [65]), breast cancer (Ayer et al. [8]), and cardiovascular disease (Mason and Denton [51]), to name a few examples. This increasing trend in the development of MDPs for medical decision making signals a broadening of scope from descriptive/predictive models to include (prescriptive) optimization models.

Chronic diseases involve a sequential progression of decisions to control or halt the course of the disease. The dependency among treatment decisions over time links the overall decision-making process, making it important to consider a holistic policy for decision making, rather than myopic decisions made in isolation at fixed points in time. The dynamic nature of decisions and a large number of possible policies motivate the need for MDPs that can be used to find optimal policies that account for uncertainty in future disease progression and health outcomes. Validated stochastic models are the foundation for sequential decision making, but creating such models is often challenging because of the imperfect nature of the longitudinal data that are commonly available.

The longitudinal data needed for building and validating stochastic models reside largely in observational data sources, such as electronic health records, insurance claims databases, laboratory data storage systems, and other forms of data that are collected routinely as part of the healthcare delivery process. These types of data are collected nonuniformly at time points when patients see their providers, fill their prescriptions, or have laboratory tests, to name a few examples. Moreover, these data are influenced by the very decisions made to treat diseases, which in turn are influenced by sources of confounding that can lead to false causal claims. For example, patients treated with blood pressure medication tend to have high blood pressure. Thus, a naïve approach that compares patients on blood pressure medication with patients who are not on blood pressure medication could wrongly conclude that blood pressure medication causes high blood pressure. Many such pitfalls exist in the use of observational data to create data-driven models for chronic diseases.

Overcoming the above challenges of building stochastic models with observational data would eliminate a major barrier to optimizing decision making but, at the same time, reveals some additional challenges that must be considered. First, the combination of uncertainty, multiple risk factors, and a large number of potential interventions makes finding optimal policies difficult because of the *curse of dimensionality*, which is a fundamental challenge of working with the multidimensional data generated for chronic diseases. Second, the imperfect estimates of model parameters, as a result of natural statistical variation or other sources of uncertainty in model estimates, raise questions about how well an "optimal" policy derived from a model with a fixed set of parameter estimates will work in practice.

The challenges of using observational data are common to many contexts, but the sequential decision-making process for chronic diseases makes these problems particularly difficult to deal with because of the need to consider how decisions made today influence future decisions and health outcomes. Methodological approaches for these types of problems include MDPs, partially observable MDPs (POMDPs), and many other related methods. The main goal of this tutorial is to provide a starting point for learning how to initiate research in the area of sequential decision making for chronic diseases using these approaches. As such, this tutorial provides guidance on model formulations, methods for fitting stochastic models for chronic diseases using longitudinal data, and the subsequent solution of these models to find policies for medical decision making.

In this tutorial, we will make some simplifying assumptions about the models that we cover. First, we assume that there is a finite set of health states and health intervention decisions. This assumption is a reasonable starting point because these models are a stepping stone to more complex models and because many diseases have a clinically meaningful discrete state definition. Second, we focus on finite-horizon nonstationary models because chronic diseases play out over the finite (but uncertain) lifetime of a patient and because they are often best modeled as nonstationary because age is an important risk factor for chronic diseases. Finally, we emphasize MDPs and POMDPs; however, we provide guidance on other approaches that have been applied to sequential decision making for chronic diseases such as *robust optimization* and *reinforcement learning*. We cannot cover all of these topics in their entirety in this short tutorial, so we provide references for the reader to learn more about specific topics along the way. Excellent resources for general reading on the topic of sequential decision making include Puterman [59], Bertsekas [8], and Sutton and Barto [72]. Sources for additional background on MDPs for medical decision making include Schaefer [63], Alagoz et al. [1], and Steimle and Denton [70]. A valuable source on "best practices" for estimating state transition models, including Markov chains, can be found in Siebert et al. [66]. Finally, Gold et al. [34] is an excellent reference on cost effectiveness in the context of medical decision making.

The remainder of this tutorial is organized as follows. First, we begin with some background on chronic diseases to set the foundation for the applications discussed in this tutorial. In Section 3, we describe the main elements of models for sequential decision making with some examples of choices that one must make during the model formulation process. Next, we provide a generic formulation of an MDP model for optimizing the time to initiate health interventions over the course of a disease. We also summarize the formulation of POMDP models because these are highly relevant to chronic diseases. In Section 4, we discuss the challenges of parameter estimation of models using longitudinal data, and we give examples in the context of diabetes treatment and cancer surveillance. In Section 5, we discuss typical data sources for model building and approaches for addressing model uncertainty and ambiguity. In Section 6, we discuss examples of alternative approaches to sequential decision making. Finally, in Section 7, we make some concluding remarks and comment on future opportunities for research on sequential decision making for chronic diseases.

2. Chronic Diseases

Chronic diseases are the leading cause of death and the largest source of cost to the U.S. health system, accounting for more than 80% of the nation's healthcare costs (Centers for Disease Control and Prevention [28]). Many patients can live for years or even decades with chronic diseases if they are detected early and treated optimally. But the decisions faced by patients and physicians are difficult because of a large number of potential interventions and the stochastic nature of diseases. The stages of the decision-making process for chronic diseases can be broadly categorized as *prevention*, *diagnosis*, *treatment*, and *posttreatment*. One reason for the difficulty in decision making for chronic diseases is that these stages are often linked because long-term measures of health, such as quality-adjusted life span, are influenced by the decisions made during all stages of the disease. A thorough understanding of risk factors and how they change as patients age is crucial for optimizing decisions in a way that balances the benefits and harms of health interventions. The benefit-versus-harm trade-off depends on (a) the risk of adverse outcomes from surgery versus the potential health benefits if the surgery is successful, (b) the reduction in the risk of future adverse events from medications versus the daily side effects from treatment, and (c) or the anxiety associated with biomarker tests versus the benefit of early detection of a disease. The presence of other diseases or conditions (comorbidities) plays an important role in weighing the benefit-versus-harm trade-off for health interventions because comorbidities can constrain the use of health interventions or

reduce the benefit of health interventions in some cases. In the remainder of this section, we briefly summarize each of the decision-making stages.

Prevention aims to reduce the incidence of disease in a healthy population by using *biomarkers* to identify patients at risk of developing a disease or in early stages of the onset of a disease when the natural course of the disease can be controlled or halted. In this tutorial, we use the general term "biomarker" to refer to the wide range of measurable indicators of disease. Biomarker tests include physiological measures associated with a condition or disease (e.g., blood pressure, heart rate, cholesterol level) and genes or proteins that can be detected using blood tests, urine tests, or tissue-based analysis. Prevention frequently involves *screening* of a healthy population using biomarkers to spot early signs of disease. Optimal decisions about biomarkers, such as the choice of which biomarker to use and the frequency of testing, can be difficult because most biomarkers have significant *false-positive* and *false-negative* rates. These errors can cause harm to healthy patients because of subsequent referrals for unnecessary diagnostic tests (false negatives) and failure to identify patients who are experiencing the onset of chronic disease (false positives).

Diagnosis aims to characterize the presence and nature of a condition or a disease for patients believed to be at risk, often using diagnostic tests and patient-reported symptoms. Diagnostic tests include imaging tests (e.g., computer tomography (CT), magnetic resonance imaging (MRI), ultrasound) and procedures such as cardiac catheterization, colonoscopy, or upper endoscopy. Similar to biomarkers, all diagnostic tests have some probability of a false-positive or false-negative outcome. Therefore, in some cases, a collection of diagnostic tests may be used in combination to make a diagnosis. Often a noninvasive test is used to decide whether to recommend a patient for a more invasive test or procedure. For example, the fecal occult blood test may be used to decide whether to perform a colonoscopy when screening for colon cancer.

Treatment frequently aims to control the risk of adverse complications and ease the burden of the disease. This long-term phase of care may also include regular *surveillance* tests to monitor the potential progression or recurrence of the disease over time. For example, patients with type 2 diabetes are recommended to have an HbA1c test—a test for monitoring long-term glucose exposure—every three months to evaluate the need for additional treatment. Treatment of chronic diseases may include prescription medications, procedures, surgery, or other health services. In the case of prescription medication, often decisions must be made about which medications to recommend. For example, there are many medications available to control the risk of cardiovascular disease, and the best choice may depend on the patient's overall risk profile and the patient's response to medication, which is unknown at the time of selection of treatment. It is important to consider both the positive benefits (e.g., cholesterol or blood pressure reduction) and the side effects (e.g., nausea, muscle pain, deterioration in liver function) of medication. In some cases, multiple health interventions may be used in combination as the disease progresses over time and a patient's risk increases. When tests or procedures are used to monitor a chronic disease, decisions about the frequency and intensity of the interventions may arise. For example, Avastin and Lucentis are two drugs that are used to control macular degeneration, a leading cause of blindness in adults. The drugs are injected in the eye in a short outpatient procedure, and the optimal frequency of injections depends on the balance between the benefits of the injections in controlling the degeneration of vision and the side effects and cost of regular injections.

Posttreatment care involves the clinical management of side effects of major treatment such as surgery or radiation therapy, sometimes requiring additional procedures, further treatment, or continual monitoring (often referred to as *surveillance*). For example, patients with localized breast cancer may have a *lumpectomy* to remove a tumor surgically. This is often followed by adjuvant radiation therapy or chemotherapy for a set course of treatment over days or weeks; moreover, it may be recommended that a patient with a history of breast cancer have more frequent mammograms (e.g., every six months instead of annually). In many cases,

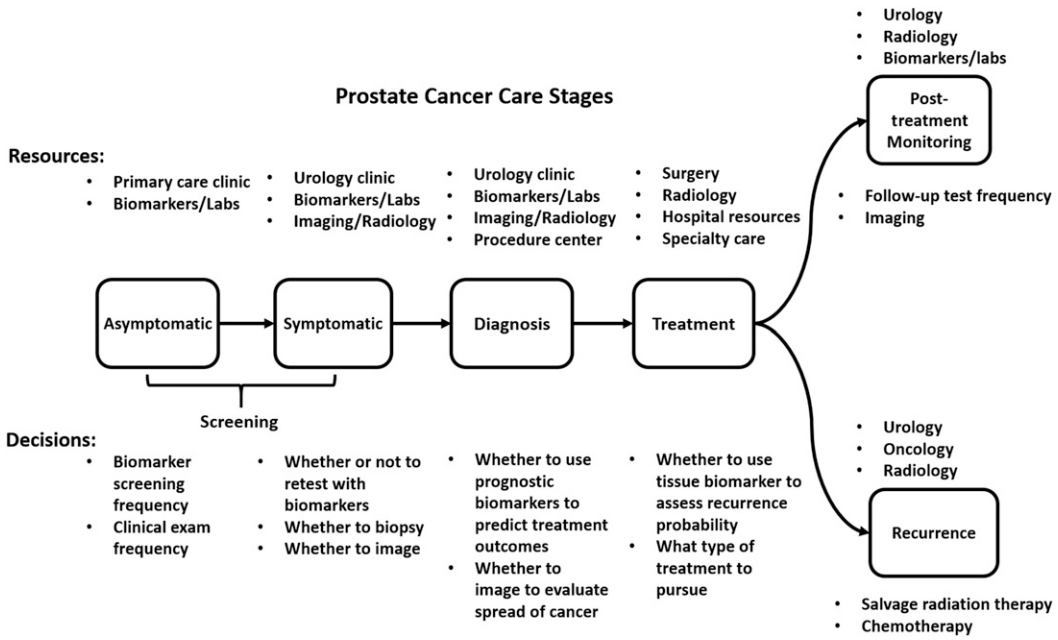
patients received regular surveillance tests with the goal of identifying possible recurrence of the disease. These could include clinical exams, biomarker tests, or procedures such as endoscopy. Patients who experience progression of chronic diseases to a terminal stage may seek *palliative care*, which is aimed at controlling the symptoms and stress of serious illness. In the worst case, the patient may progress to *hospice care*, which concentrates on symptom care at the end of life.

The above categories of decisions—prevention, diagnosis, treatment, and posttreatment care—are often interrelated as a result of the propensity for future (anticipated) decisions to influence what is best at present. One such example is in the area of prostate cancer, which is a disease with well-defined stages, each with important and unique decisions as illustrated in Figure 1. Prostate cancer screening is commonly implemented using the prostate-specific antigen (PSA) test. This simple blood test is used to detect latent prostate cancer, before it becomes symptomatic when surgery is a potential cure. However, the risk of mortality for some prostate cancers is very low, and therefore *other-cause mortality risk*—that is, the risk of dying from any cause other than prostate cancer—plays an important role in determining the benefit to a patient of detecting prostate cancer. For men over the age of 75, the American Urology Association recommends against PSA screening because older patients are unlikely to be treated because the benefit of treatment is very low compared with the risk of "other-cause" mortality (e.g., cardiovascular disease or any cause of death other than prostate cancer). This interdependence between disease stages is very common for chronic diseases, which motivates the importance of a sequential decision-making approach.

3. Markov Decision Processes for Chronic Diseases

We begin by describing the standard elements of an MDP for chronic disease including decision epochs, the time horizon, a finite set of health states, a finite set of decisions about health interventions, state transition probabilities, and a reward function. We introduce generic

Figure 1. An example of the stages of care and the decisions and resources at each stage for prostate cancer beginning with a screening of a healthy population through diagnosis, treatment, and post-treatment monitoring and recurrence.



definitions throughout this section to develop a model that could fit many contexts. In Section 4, we provide specific examples in the context of treatment for diabetes and surveillance of prostate cancer to illustrate the use of these models for optimizing medical decision making. The following are the standard elements of an MDP formulation:

- *Decision epochs*: Decisions are made at each epoch in a discrete set of decision epochs, which are fixed and predetermined time points over the finite horizon at which decisions are made. The selected time interval between decision epochs for chronic disease should be at least as frequent as a typical clinical decision occurs in practice. In the case of type 2 diabetes—an example considered later in this tutorial—decisions about which medications to initiate are made every six months as recommended by the American Diabetes Association [4]. If the decisions themselves include whether and when to perform tests, to collect additional information, then more frequent intervals may be appropriate to avoid biasing the decisions toward longer intervals.
- *Time horizon*: The time horizon in an MDP may be finite or infinite. An infinite time horizon must include a discount factor on future rewards to guarantee the total rewards are bounded. A discount factor may also be used in a finite horizon MDP, but it is not a requirement for a well-formulated model. Although the disease life cycle progresses over a finite horizon, some researchers elect an infinite-horizon approach when the time between decision epochs is short relative to the length of the time horizon. Another deciding factor for choosing an infinite-horizon model is whether stationary transition probabilities and rewards are reasonable assumptions. For example, some diseases are highly nonstationary because age is a risk factor (e.g., cardiovascular disease risk increases exponentially over time), whereas others may be reasonably approximated by a stationary MDP.
- *States*: At each decision epoch, the system described by an MDP model is in a certain state. The choice of states is one of the most important decisions when formulating an MDP model. The choice should be based on the minimal information required for clinical decision making at a given decision epoch. The specific definition depends on the particular disease and may include risk factors defining the patient's health status, demographic information, and the relevant medical history. For example, a model for prevention of cardiovascular disease might include cholesterol, blood pressure, and other factors relevant to predicting cardiovascular disease outcomes as established in the medical literature (e.g., gender, smoking status, previously initiated medications). In some cases, the dimensionality of the state may be reduced by using an established aggregate risk score. For example, models for liver transplant decisions have used the discrete Mayo End-stage Liver Disease (MELD) score, which is a single score based on multiple risk factors for liver disease (Alagoz et al. [2]). In contexts where the state is defined by a continuous risk factor (e.g., cholesterol, blood pressure, blood sugar), it is common to discretize the state space to obtain a tractable approximation of the continuous model. A finer discretization may be more representative of the true continuous state space, but it also increases the size of the state space and therefore the computation required to solve the model. Furthermore, a finer discretization decreases the number of observed transitions among states in a longitudinal data set, introducing more sampling error into the estimates of the transition probabilities. Regnier and Shechter [61] provide a discussion of the trade-off between the model error, caused by coarse discretization of states, and the sampling error as a result of a finer discretization. In addition to model accuracy, it may be necessary to consider clinically relevant thresholds for defining discrete states. This consideration is important if the MDP model will be compared to clinical guidelines.
- *Decisions:* The decisions (often referred to as *actions*) in an MDP may include many types of health interventions such as oral medications (e.g., statins for cholesterol reduction), injectable medications (e.g., insulin for glucose control), and diagnostic tests for disease screening or surveillance (e.g., cluster of differentiation 4 (CD4) count for HIV). Diagnostic tests could also include molecular biomarkers implemented as blood tests, urine tests, imaging tests such as MRIs, CT scans, or other means of assessing the risk of a latent disease such as

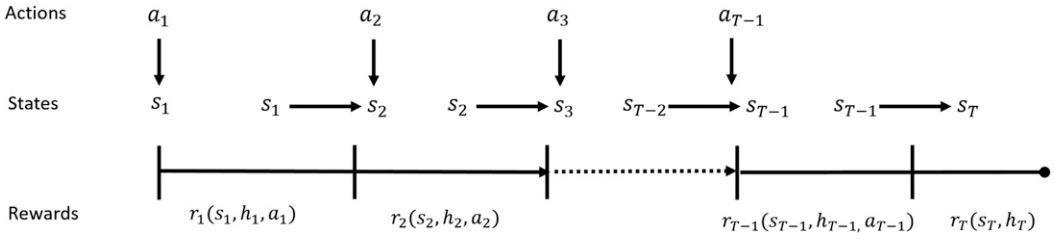
cancer. In the model we present in this section, we consider a finite set of actions that are binary decisions (e.g., initiate a nominal dose of medication, refer a patient for an MRI, discontinue treatment). Continuous decisions do arise and may be appropriate in some cases (e.g., selecting a medication dose), but often they are reasonably modeled as a discrete decision because there are typically a discrete set of options that are commonly used in practice (e.g., low, medium, or high dose of medication). The term *policy* is used to define a mapping of MDP states to actions. Thus, the goal of solving an MDP is one of finding the optimal policy, which we discuss in Section 3.1.

- *Transition probabilities:* Conditional probabilities define the random change in state from one decision epoch to the next. Under the Markov assumption of an MDP, the probability of transitioning to a given state in the next decision epoch depends only on the current state. The transition probabilities describing the progression of the disease constitute a model known as a *natural history model* of the disease. Creating such models is challenging because medical records contain data about patients who have been treated. Therefore, to estimate transition probabilities among "untreated" states, it is necessary to transform the longitudinal data by removing the estimated effect of treatment on the risk factors (e.g., oral medications for diabetes such as metformin, lower blood sugar, change in the natural course of the disease).
- *Rewards:* At each decision epoch, a reward is received that may depend on the state, action, and decision epoch (the term "reward" is commonly used, but it may also refer to a cost or penalty of some time in the context of a minimization problem). The rewards in a chronic disease MDP model may be associated with health or economic implications (e.g., costs). The specific choice of reward may differ depending on whether the decision maker is a patient, physician, or third-party payer (e.g., BlueCross BlueShield, Medicare). Health interventions are intended to offer some reward to the patient, such as a potentially longer life or improved quality of life. But these benefits come at a "cost" to the patient, whether it is a reduction in quality of life, caused by the side effects of a health intervention, or a financial cost such as medication co-pays or hospital expenses. Health services researchers typically use quality-adjusted life years (QALYs) to quantify the quality of a year of life. A QALY of 1 represents a patient in perfect health with no adverse impact from the disease and no side effects as a result of health interventions. As the patient's quality of life decreases—whether from health intervention side effects or disablement from a disease—the patient's expected QALYs will decrease. The adjustments for reduced health as a result of symptoms of a disease or side effects of interventions are called *disutilities*, which are numeric estimates of harm to quality of life on a scale of 0 to 1. They are often estimated by survey studies that elicit patient estimates of harm associated with health outcomes using survey methods (see Torrance [74] for a review of standard methods including *standard gamble* and *time trade-off*). Some MDP models are only concerned with maximizing a patient's QALYs. Other models take a societal perspective and attempt to balance the health benefits of treatment with the corresponding monetary costs of health interventions. A common approach to balance competing objectives uses a *willingness-to-pay* factor, which assigns a monetary value to a QALY. In this case all the rewards are in dollars, and the *net monetary benefit* (NMB) is an often-used criterion that is the difference between the reward for QALYs and the cost for health interventions. Commonly used values for the willingness-to-pay factor are \$50,000 and \$100,000 per QALY, but the most appropriate value to use is often debated (Rascati [60]).

3.1. MDP Model Formulation

In this section, we give a generic mathematical formulation of an MDP for a chronic disease. This formulation is intended to convey a general conceptual understanding of MPDs for chronic diseases, but any specific application is likely to require some modifications to tailor the model. Figure 2 illustrates the sequential decision-making process over a finite horizon using the notation defined in this section. Decisions are revisited periodically over the set of decision epochs: $\mathcal{T} \equiv \{1, 2, \ldots, T\}$. The disease states are a combination of *disease status* and

Figure 2. Illustration of the sequential decision-making process for an MDP, including actions, state transitions, and rewards, from the start to end of the finite-time horizon.



intervention history. The intervention history is included as part of the state definition because health interventions often have long-term, lasting effects and thus must be incorporated into the state definition to retain the Markov property. The set of disease status states is $\mathcal{S} \equiv \{S_1, S_2, \ldots, S_{|\mathcal{S}|}\}$, and the complete set of possible health intervention histories at epoch t is $\mathcal{H}_t \equiv \{H_1, H_2, \ldots, H_{|\mathcal{H}_t|}\}$. The disease status set and the health intervention history set are indexed at each epoch t by $s_t \in \mathcal{S}$ and $h_t \in \mathcal{H}_t$, respectively. Most MDP models also have at least one *absorbing state*, $\mathcal{D}$, which we assume is defined in the set of health status states, $\mathcal{S}$. Depending on the disease context, $\mathcal{D}$ could represent major complications of the disease, death, or some other cause of departure from the decision process (e.g., organ transplant).

At each decision epoch t, the action, a_t, is selected from a set of available interventions, $\mathcal{A}_t(s_t, h_t)$, that may depend on the patient's health status, s_t, and health intervention history, h_t. There are many possible reasons that the current action a_t would depend on the patient's health or intervention history. For example, as a patient's health status deteriorates, certain interventions may be too risky or may be unlikely to yield a benefit to a patient (e.g., a patient with metastatic cancer may not benefit from surgery to remove a tumor). There may also be conflicts between certain interventions (e.g., dangerous interactions between prescribed medications). In some cases, there may be a clinical reason why a certain order or partial order of health interventions is appropriate (e.g., less invasive tests are often used before more invasive tests). Because of the potential dependency between past interventions and current actions, the set of health intervention histories is updated at each epoch via the following set operation: $\mathcal{H}_t \equiv \mathcal{H}_{t-1} \times \{a_t\}$. Because of the constraints on interventions, it is sometimes the case that the set of available interventions $\mathcal{A}_t(s_t, h_t)$ is nonincreasing over time: $\mathcal{A}_1(s_1, h_1) \supseteq \mathcal{A}_2(s_2, h_2) \supseteq \cdots \supseteq \mathcal{A}_{T-1}(s_{T-1}, h_{T-1})$. This relationship reflects the fact that the available options tend to decrease over time as a patient ages and progresses to later stages of a disease.

At each decision epoch t, and for each state pair (s_t, h_t), the decision maker selects an action $a_t \in \mathcal{A}(s_t, h_t)$ and receives a reward of $r_t(s_t, h_t, a_t)$. On the basis of the above definition of states and actions, there are two types of probabilities in this MDP: (1) transition probabilities among (transient) disease and health intervention states and (2) transitions to the absorbing state(s). The complete set of transition probabilities is summarized in the following equation:

$$p_t(s_{t+1}|s_t, h_t, a_t) = \begin{cases} (1 - p_t(\mathcal{D}|s_t, h_t, a_t)) \cdot p_t(s_{t+1}|s_t, h_t, a_t) & \text{if } s_t,\ s_{t+1} \in \mathcal{S}\backslash\{\mathcal{D}\}, h_t \in \mathcal{H}_t \\ p_t(\mathcal{D}|s_t, h_t, a_t) & \text{if } s_{t+1} = \mathcal{D} \text{ and } s_t \in \mathcal{S}\backslash\{\mathcal{D}\}, h_t \in \mathcal{H}_t \\ 1 & \text{if } s_t = s_{t+1} = \mathcal{D}, h_t \in \mathcal{H}_t \\ 0 & \text{otherwise,} \end{cases}$$

where $p_t(\mathcal{D}|s_t, h_t, a_t)$ is the probability of entering the absorbing state. This definition implicitly assumes that transitions to the absorbing state, $\mathcal{D}$, and the transient states, $\mathcal{S} \times \mathcal{H}_{t+1}$, at epoch $t + 1$ are conditionally independent; that is, their dependence is described entirely by the current state pair (s_t, h_t) and action a_t.

From the above definitions, the goal is to find a policy that maximizes the expected total discounted rewards over the time horizon, as follows:

$$\max_{\pi\in\Pi}\Big\{\mathbb{E}_{\pi}\Big[\sum_{t=1}^{T-1}\lambda^{t-1}r_t(s_t,h_t,\pi(s_t,h_t))+\lambda^{T-1}r_T(s_T)\Big]\Big\}, \tag{1}$$

where the expectation is taken with respect to the stochastic process induced by policy π, which is a vector of *decision rules*, which in turn are vectors of optimal actions for each epoch t. We denote a decision rule by vector $\vec{d}_t^{\,*}\equiv(a_t^*(s_1,h_1),\ldots,a_t^*(s_{|\mathcal{S}|},h_{|\mathcal{H}|}))$. The set of optimal decision rules at each epoch defines the optimal policy $\vec{\pi}^*\equiv(d_0^*,d_1^*,\ldots,d_{T-1}^*)$. The set of all possible policies is denoted by Π in Equation (1). The optimal policy defines the complete set of actions for every possible decision epoch and health state combination.

The optimal policy for the problem defined by Equation (1) can be found by solving the *optimality equations* for a stochastic dynamic program (also known as Bellman's equations), which are written as follows:

$$v_t(s_t,h_t)=\max_{a_t\in\mathcal{A}(s_t,h_t)}\Big\{r_t(s_t,h_t,a_t)+\lambda\sum_{\forall s_{t+1}\in\mathcal{S}}p(s_{t+1}|s_t,a_t)v_{t+1}(s_{t+1},h_{t+1})\Big\}$$
$$t\in\mathcal{T}\backslash\{T\},\ s_t\in\mathcal{S},\ h_t\in\mathcal{H}_t$$
$$v_T(s_T,h_T)=r_T(s_T,h_T),\quad s_T\in\mathcal{S},\ h_T\in\mathcal{H}_T,$$

where $v_t(s_t,h_t)$ is the expected *value to go* if the optimal action is followed in each decision epoch starting at epoch t, when the patient is in health status state s_t and intervention history h_t, and future rewards are discounted by $\lambda\in(0,1]$. The fact that the above optimality equations provide an optimal policy is easily proven by induction (see section 4.3 of Puterman [59] for a proof). Discounting is common in MDPs to account for the time value of rewards. In healthcare studies, an annual discount factor of 0.97 is commonly used when the criterion is a monetary cost to account for the time value of money. Discounting of QALYs is also common in the context of cost-effectiveness analysis that considers the ratio of the change in cost to the change in QALYs for a health intervention (known as the *incremental cost-effectiveness ratio*). A discussion of discount factors can be found in Gold [34]. The value function at epoch T is a boundary condition determined by the end-of-horizon reward, $r_T(s_T,h_T)$.

The above MDP formulation has similarities to many MDPs proposed in the literature on medical decision making for chronic diseases, but there are also some notable extensions that have received attention. For example, *semi-Markov decision processes* allow for uncertainty in the time between state transitions by employing a continuous time approach that allows for a probability distribution over the amount of time spent in a particular state (Serfozo [64]). Chou et al. [18] provide an example of the use of a semi-Markov process for optimizing the time to initiate medical treatment. *Factored MDPs* recognize that some problems have multiple independent variables that define the state space, a characteristic that can be exploited to some computational advantage (Degris and Sigaud [23]). Another significant extension is to consider the addition of constraints (e.g., constraints on total cost over multiple decision epochs), which leads to challenges in developing algorithms because many such problems no longer retain the attractive decomposable structure of a dynamic program (Altman [3]). Finally, the above formulation assumes that the decision maker is *risk neutral*. Approaches to generalize MDPs to consider risk aversion include the use of utility functions (Howard and Matheson [39]), percentile risk measures (Filar et al. [28]), and more recent risk measures such as conditional value at risk (Chow and Ghavamzadeh [19]).

3.2. Partially Observable MDP Model Formulations

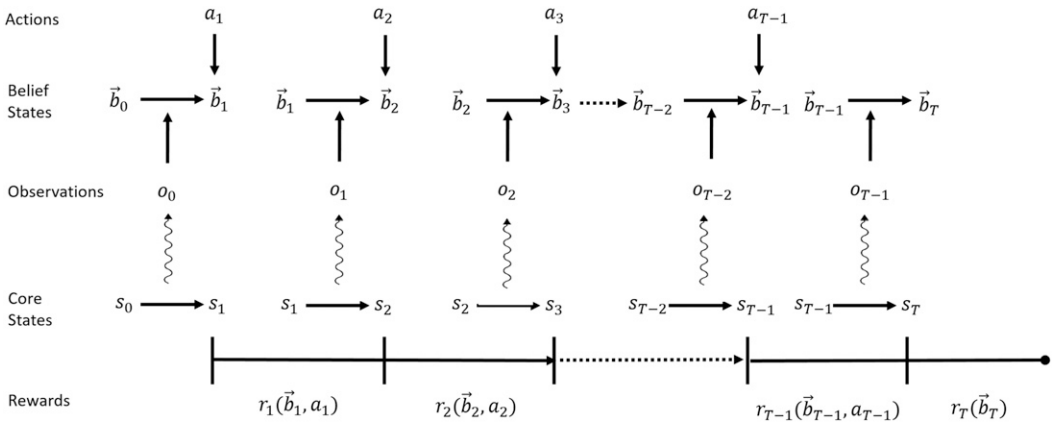
POMDPs extend MDPs to contexts in which perfect information about the health state of the patient is not available. We emphasize this particular extension among the many alternatives

because it applies to the many diseases that have an asymptomatic latent period (e.g., many cancers go undetected because of an absence of symptoms). Examples of the use of POMDPs for chronic diseases include breast cancer (Ayer et al. [78], Maillart [49]), colorectal cancer (Erenay et al. [27]), prostate cancer (Zhang et al. [81]), and heart disease (Hauskrecht and Fraser [36]), to name a few. For these types of diseases, screening and diagnostic tests often provide useful information, but false-positive and false-negative test results prevent the true health state from being known with certainty. POMDPs assume that the decision maker does not know the exact health state of the patient. Instead, the health state is replaced by a *belief state* that defines a probability distribution over the finite set of disease states.

Next, we describe the most important elements of POMDPs. To the extent possible, we use the same notation as the previous sections. For a more thorough description of POMDPS in general, the reader is referred to reviews by Monahan [52] and Lovejoy [48]. A tutorial by Cassandra provides an excellent nontechnical introduction to POMDPs (`POMDP.org`) as well as references to more recently developed methods.

- *Core states and observations:* In a POMDP, the state space is defined by a set of core states (also known as *latent states* or *hidden states*) and an *observation process* (also referred to as a *message process* or an *emission process*). Figure 3 illustrates this sequential state transition, observation, and decision process. For chronic diseases, the core states correspond to the true health of a patient, such as is cancer-free, has noninvasive cancer, has invasive cancer, or is in treatment. As in the previous section, we let s_t index the health status states in the set $\mathcal{S}$ (note that in this section, we suppress consideration of intervention histories, $\mathcal{H}_1, \ldots, \mathcal{H}_T$, for simplicity). To a clinician, some of these states are not directly observable, so the true health state of the patient is not known with certainty. The observation process corresponds to observable test results (e.g., a mammogram for breast cancer, a fecal occult blood test for colorectal cancer, imaging for retinopathy). The core state process and the observation process are tied together probabilistically through an *information matrix*. Each row of the information matrix corresponds to a core state, and each column entry is the probability of a particular observation conditional on the core state. The relationship between the core and observation processes and the observed test results can be used to estimate the belief vector sequentially via Bayesian updating.
- *Decisions:* To decide on the action set for a POMDP, one must identify which screening or treatment options to consider. In the context of POMDPs, it is often the case that actions involved the choice of whether and when to use screening or diagnostic tests that

Figure 3. Illustration of the sequential decision-making process for a POMDP, including actions, observations o_t, core state transitions, and rewards, from the start to the end of the finite-time horizon. Observations are emitted from the system before the state transition in each decision epoch. The observations and prior belief $\vec{b}_t$ are used to update the belief vector $\vec{b}_{t+1}$. Actions are based on the belief vector at each decision epoch.



provide information about the likelihood the patient is in a particular health state. Decisions about which actions to consider also have implications on the computational difficulty because as the number of actions increases, the computational difficulty of finding optimal policies increases exponentially (Monohan [52]).

- *Bayesian updating and optimality equations:* At each decision epoch t, an action is selected. The observations that follow inform the next choice of action at epoch $t+1$ through Bayesian updating of the *belief vector*. The belief vector at epoch t is denoted by $\vec{b}_t$, and it has elements, b_{t,s_t}, that denote the probability that the patient is in core state $s_t \in \mathcal{S}$ at epoch t. The *information matrix* has elements, denoted by $q_t(o_t|s_t, a_t)$, which define the probability of observing outcome $o_t \in \mathbb{O}$, where $\mathbb{O}$ is a finite set of possible observations (e.g., biomarker test results based on a discrete set of clinically relevant ranges), given the core state of the patient is s_t and action a_t was chosen. The belief vector is updated via Bayesian updating at the start of epoch $t+1$, immediately after observing o_t at the end of epoch t, using Bayes' rule as follows:

$$b_{t+1,s_{t+1}} = \frac{\sum_{s_t \in \mathcal{S}} b_{t,s_t} p_t(s_{t+1}|s_t, a_t) q_t(o_t|s_t, a_t)}{\sum_{s_t, s_{t+1} \in \mathcal{S}} b_{t,s_t} p_t(s_{t+1}|s_t, a_t) q_t(o_t|s_t, a_t)}, \tag{2}$$

where the numerator is the probability of transition to state s_{t+1} and observing o_t, and the denominator is the probability of observing o_t taken over all possible states to which the patient may have transitioned.

- *Rewards:* At each decision epoch, t, a reward is received that depends on the current information, including the core state, observation, and action. In this partially observable context, there are multiple alternatives for defining this dependency. We choose the following form of the rewards:

$$r_t(\vec{b}_t, a_t) = \sum_{s_t \in \mathcal{S}} b_{t,s_t} r_t(s_t, a_t), \tag{3}$$

which is the expectation of rewards defined on the core state s_t and action a_t. Given the above definitions, the resulting optimality equations are as follows:

$$v_t(\vec{b}_t) = \max_{a_t \in \mathcal{A}(\vec{b}_t)} \left\{ r_t(\vec{b}_t, a_t) + \lambda \sum_{s_t \in \mathcal{S}, s_{t+1} \in \mathcal{S}, o_{t+1} \in \mathbb{O}} b_{t,s_t} p_t(s_{t+1}|s_t, a_t) q_{t+1}(o_{t+1}|s_t, a_t) v_{t+1}(\vec{b}_{t+1}) \right\}$$

for $t \in \mathcal{T} \setminus \{T\}$ and a terminal reward vector $v_T(\vec{b}_T) = r_T(\vec{b}_T)$, where $r_T(\vec{b}_T)$ is the expected reward to go beyond the end of the time horizon as a function of the belief state. The equations look similar to the standard MDP described in Section 3.1; however, approaches for solving these problems, as we shall see in Section 3.3, are quite different because $\vec{b}_t$ is continuous.

In a POMDP model, the decision maker can take actions to gain information about the state of the system. For example, the problem of optimizing screening decisions is modeled as a POMDP where the actions at each epoch are the different choices of screening tests available, including the option not to screen. Performing a screening test may not change the natural progression of the disease, but it can provide the decision maker with valuable information about the true health state of the patient, which in turn may be used to decide whether to do more invasive testing such as biopsy or radiologic imaging. Many POMDPs used in medical applications deal with decisions about whether and when to collect information to learn about the health status of patients over time.

3.3. MDP and POMDP Solution Methods

The appropriate method for solving an MDP depends on whether it is an infinite-horizon or finite-horizon model and the size of the state and action spaces. Methods such as *policy iteration*, *value iteration*, and *linear programming* have been used to solve infinite-horizon

problems, whereas *backward induction* is typically used to solve finite-horizon problems based on an end-of-horizon boundary condition (see Algorithm 1 for a pseudocode description of this method). The general references given at the end of Section 1 describe these methods in detail.

Algorithm 1 (Backward Induction Algorithm for Finite-Horizon MDP)

1: **Input:** MDP data elements: decision epochs, states, actions, transition probability matrix, rewards, discount factor
2: **Boundary Condition:** $v_T(s_T, h_T) = r_T(s_T, h_T)$, for all $s_T \in \mathcal{S}, h_T \in \mathcal{H}_T$
3: **Backward Induction:**
4: **for** $t = T-1$ to 1 **do**
5: **for** all $s_t \in \mathcal{S}$ and $h_t \in \mathcal{H}_t$ **do**

$$v_t(s_t,h_t) = \max_{a_t \in \mathcal{A}_t}\{r_t(s_t, h_t, a_t) + \lambda \sum_{s \in \mathcal{S}} p_t(s_{t+1} \mid s_t, h_t.a_t) v_{t+1}(s_{t+1}, h_{t+1})\}$$

$$a_t^*(s_t,h_t) = \arg\max_{a_t \in \mathcal{A}_t}\{r_t(s_t, h_t, a_t) + \lambda \sum_{s \in \mathcal{S}} p_t(s_{t+1} \mid s_t, h_t.a_t) v_{t+1}(s_{t+1}, h_{t+1})\}$$

6: **end for**
7: **end for**
8: **Return:** Optimal Policy

A common problem with practical MDP formulations is that they are subject to the curse of dimensionality because the size of the state space grows exponentially with the number of health risk factors defining the patient's health state over time. Approximation algorithms can be used to circumvent this problem. There has been a large amount of research on approximate dynamic programming in recent years. These approaches tend to be highly context dependent, and with a few notable exceptions, very little work has been done in the context of chronic diseases. One example of the use of approximate dynamic programming arises in the context of treatment decisions for infertility (He et al. [37]). Books by Bertsekas [9] and Powell [58] provide a thorough review of approximation methods for MDPs.

Many MDP models for chronic diseases have structural properties that can help explain the optimal policies obtained from solving MDPs and, in some cases, can be exploited for computational gains. One such property is the *increasing failure rate* (IFR) property of transition probability matrices. In the context of chronic diseases, the IFR property means that there is an ordering of states (e.g., least to most healthy) such that the worse the health status of the patient is, the more likely that the health status will become even worse. Mathematically, it is defined as follows:

Definition. A transition probability matrix at epoch t has the IFR property if $\sum_{s_{t+1}=k}^{|\mathcal{S}|} p_t(s_{t+1}|s_t, a_t)$ is nondecreasing in s_t for all $k \in \mathcal{S}$ and $a_t \in \mathcal{A}$.

Usually, the state ordering naturally follows some measure of the severity of the chronic disease (e.g., low to high risk of a disease complication). For certain problems, the IFR property together with some additional (and generally nonrestrictive) conditions guarantee an optimal threshold policy (see chapter 4 of Puterman [59] for a thorough discussion of this topic). These conditions have been used, for example, in the context of HIV (Shechter et al. [65]), liver disease (Alagoz et al. [2]), and type 2 diabetes (Kurt [44]) to prove the existence of an optimal *control-limit policy* under an ordering of states. A control-limit policy is one in which one action is optimal for all states below a certain threshold state (e.g., wait to transplant if the MELD score is below 25) and another action is optimal for all states at or above a certain value (e.g., transplant if the MELD score is at or above 25). In addition to providing insight into optimal policies for MDPs, proving the existence of a control-limit policy can also decrease the computational effort required to solve an MDP model because the value function does not need to be explicitly calculated for every state–action pair.

POMDPs are generally much more challenging to solve than MDPs. Whereas MDPs can be solved in polynomial time, POMDPs have time complexity that grows exponentially in the number of observations and actions; moreover, the dimension of the belief vector

increases with the number of core states. Early methodological studies of POMDPs focused on exact methods that exploit the fact that the optimal value function for a POMDP is convex, and in the finite-horizon case, it is piecewise linear and expressible using a finite set of supporting hyperplanes known as *α-vectors* (Smallwood and Sondik [68]). The first exact method was the *single-pass* method, given in Algorithm 2, which, similar to Algorithm 1, uses backward recursion starting with a boundary condition in epoch T. In contrast to Algorithm 1, however, Algorithm 2 exploits the following piecewise linear convex property of the optimal value function:

$$v_t(\vec{b}_t) = \max_{\vec{\alpha}_t \in \Omega_t} \left\{ \vec{b}_t \cdot \vec{\alpha}_t \right\}, \tag{4}$$

where Ω_t is a finite set of $|\mathcal{S}|$-dimensional α-vectors. Each α-vector has a corresponding action; therefore, Equation (4) encodes the optimal action at each epoch t and belief point $\vec{b}_t$. At epoch t, Ω_t can be recursively generated using the optimality equations:

$$v_t(\vec{b}_t) = \max_{a_t \in \mathcal{A}_t} \left\{ \vec{b}_t \cdot \vec{r}_t(a_t) + \lambda \sum_{o_t \in \mathbb{O}} \max_{\vec{\alpha}_{t+1} \in \Omega_{t+1}} \left\{ \vec{b}_{t+1} \cdot \vec{\alpha}_{t+1} \right\} \bar{p}_t(o_t | \vec{b}_t, a_t) \right\}, \tag{5}$$

where the first term in the maximization is the dot product of the rewards in Equation (3) (i.e., $\vec{r}_t(a_t)$ is an $\mathcal{S}$-dimensional vector of rewards for all states), $s_t \in \mathcal{S}$, and $\bar{p}_t(o_t \mid \vec{b}_t, a_t)$ is the probability of observing o_t given the belief state $\vec{b}_t$ and action a_t. Substituting in Equation (2), this can be rewritten as follows:

$$v_t(\vec{b}_t) = \max_{a_t \in \mathcal{A}_t} \left\{ \vec{b}_t \cdot \left(\vec{r}_t(a_t) + \lambda \sum_{o_t \in \mathbb{O}} \max_{\vec{\alpha}_{t+1} \in \Omega_{t+1}} \sum_{s_t \in \mathcal{S}} q_t(o_t | s_{t+1}, a_t) p_t(s_{t+1} | s_t, a_t) \vec{\alpha}_{t+1} \right) \right\} \tag{6}$$

for all $\vec{b}_t$ in the unit simplex and all epochs $t \in \mathcal{T} \setminus \{T\}$, and the boundary condition is

$$v_T(\vec{b}_T) = \vec{b}_T \cdot \vec{r}_T$$

for all $\vec{b}_T$ in the unit simplex. The vector $\vec{r}_T$ has elements corresponding to the end-of-horizon reward for each state $s_t \in \mathcal{S}$. The belief vector has been factored out in Equation (6) to separate the α-vector from the belief vector. Given the above properties, the problem of solving a POMDP is equivalent to finding the α-vector set that describes $v_t(\cdot)$ at each epoch $t \in \mathcal{T}$. The single-pass algorithm, mentioned previously, constructs Ω_t from the previous α-vector set, Ω_{t+1}, as follows:

$$\Omega_t \equiv \left\{ \vec{\alpha}_t | \vec{\alpha}_t = \vec{r}_t(a_t) + \lambda \sum_{o_t \in \mathbb{O}} \sum_{s_t \in \mathcal{S}} q_t(o_t | s_{t+1}, a_t) p_t(s_{t+1} | s_t, a_t) \vec{\alpha}_{t+1} \mid \forall a_t \in \mathcal{A}_t, \forall \vec{\alpha}_{t+1} \in \Omega_{t+1} \right\}.$$

As defined, Ω_t is the finite set of all possible α-vectors; however, there are typically a large number of vectors that are unnecessary because one or more other vectors dominate them; that is, there is no belief point at which the given vector is necessary to define the *epigraph* of the optimal value function. An α-vector $\alpha \in \Omega_t$ can be eliminated through a process called *pruning* that involves solving the following linear program, $LP(\alpha_t, \Omega_t)$, for every alpha vector α_t:

$$\begin{aligned} \min \quad & z = y - \sum_{s_t \in \mathcal{S}} b_{t,s_t} \alpha_t(s) \\ \text{subject to} \quad & \sum_{s_t \in \mathcal{S}} b_{t,s_t} \alpha'_t(s_t) \le y, \ \forall \vec{\alpha}'_t \in \Omega_t \setminus \{\vec{\alpha}_t\}, \\ & \sum_{s_t \in \mathcal{S}} b_{t,s_t} = 1, \\ & 0 \le b_{t,s_t} \le 1, \forall s_t \in \mathcal{S}. \end{aligned} \tag{7}$$

If the optimal solution to the linear program, $LP(\alpha_t, \Omega_t)$, in Equation (7) is such that $z^* > 0$, then $\vec{\alpha}_t$ is nondominated; otherwise, it can be removed without altering the optimal value function or policy. Algorithm 2 provides pseudocode for generating the minimal set of nondominated alpha vectors at each decision epoch from which the optimal policy for any given belief state, $\vec{b}_t$, can be computed using Equation (5).

Algorithm 2 conveys a conceptual understanding of an exact approach for POMDPs; however, it is suitable only for very small POMDPs. Many authors have built on this early approach for solving POMDPs by developing more efficient ways of pruning unnecessary α-vectors, including *incremental pruning* (Cassandra et al. [12]) and the *witness method* (Litman [46]). Even these more efficient exact methods are generally limited to small POMDPs. Thus, approximation methods have been the focus for practical POMDPs, such as those that arise in the context of chronic diseases. Perhaps the most well-known approximation method for POMDPs is the *fixed-finite-grid algorithm* proposed in Eckles [25]. This approach approximates the continuous belief space with a finite set of belief points, resulting in a completely observable MDP that approximates the POMDP. Many enhancements, including variable grid-based approaches, have built on this early idea. The reader is referred to Lovejoy [48] for a survey of approximation methods including finite-grid based approximations. A more general survey of theory and methods for solving POMDPs, including exact and approximation methods, can be found in Kaelbling et al. [41].

Algorithm 2 (Single-Pass Algorithm for Finite-Horizon POMDP)
Input: POMDP elements: decision epochs, states, actions, transition probability matrix, information matrix, rewards, discount factor
Boundary Condition: $\Omega_T = \{\vec{r}_T\}$
Backward Induction:
for $t = T-1$ to 1 **do**
  Generate Ω_t
  for all $\alpha_t \in \Omega_t$ **do**
    if $LP(\alpha_t, \Omega_t) > 0$ **then** $\Omega_t \leftarrow \Omega_t \backslash \{\alpha_t\}$
    end if
  end for
end for
Return: Minimal α-vector set Ω_t for all $t \in \mathcal{T}$.

3.4. Software for Solving MDPs and POMDPs

There are numerous software implementations of algorithms for solving MDPs and POMDPs. We provide a few examples here, although this is not intended to be a comprehensive list of all available software. For MDPs, the MDPToolbox package, which is available for many environments including MATLAB, R, and Python, provides solvers for discrete-time MDPs including backward recursion for finite-horizon problems and methods such as *value iteration* *policy iteration* for infinite-horizon problems (Chadès et al. [15]). The JuliaPOMDP package includes implementations of methods for MDPs and some algorithms for POMDPs in the Julia programming language. Poupart et al. [57] provide source code for approximations to POMDPs with optimality gaps. The tutorial on POMDPs (`POMDP.org`) provides source code for implementations of some of the more popular methods.

4. Data-Driven Model Parameterization for MDPs and POMDPs

With the appropriate definitions of models for sequential decision making established, we now discuss the issue of how to estimate model parameters using longitudinal data from electronic health records and other sources. We use two examples to illustrate methods for estimating natural history models: treatment for type 2 diabetes and surveillance of patients with prostate

cancer. For each example, we provide a detailed description of how the model parameters were estimated, and we present results for the optimal policies obtained from the models.

4.1. Model Parameterization for MDPs

The two main categories of data for MDPs are rewards and transition probabilities. The choice of reward parameters depends on the criteria to be considered, which can vary significantly depending on the decision maker's perspective. Common examples include expected life span, QALYs, the risk of a major health complication, and the cost of health services. As discussed in Section 3, QALYs refine the expected life span measure to account for the effect of disease outcomes and side effects of interventions using disutilities (also known as *utility decrements*). Disutilities are numerical estimates used to adjust a year of perfect health quantitatively as a result of the impact of disease or health intervention side effects. These estimates are often drawn from the health services research literature based on survey studies of patients to elicit disutility estimates. In many cases no disutility estimates are available, and one must rely on expert opinion or find estimates of disutilities for similar interventions. For example, the disutility associated with one procedure (e.g., cardiac catheterization) may serve as a plausible estimate for another (e.g., endoscopy). Disutilities are often included in sensitivity analysis because of the limited availability of data from which to estimate them and because the optimal policy for an MDP is frequently sensitive to the choice of these parameters.

When transition probabilities are estimated using longitudinal data, the estimates are highly dependent on the definition of the states that define the Markov chain. Small numbers of states often lead to dense transition probability matrices for which good point estimates can be obtained (e.g., categorizing blood pressure into two states, low and high); however, the small number of states means that the classification of alternative disease states is coarse, and the accuracy of the model may be poor. Alternatively, a large number of states may be used to boost model accuracy; however, this comes at the expense of a larger number of transition probabilities to be estimated, which increases the statistical error in the model parameters. In general, the choice of state must carefully weigh clinical expertise, computational considerations, and constraints caused by the limited availability of sample data for model estimation.

4.1.1. Example: Treatment for Type 2 Diabetes. We provide an example of estimation of a transition probability matrix in the context of type 2 diabetes where the states are defined by HbA1c, a commonly used biomarker for estimating long-term blood sugar exposure based on the percentage of blood cells with glucose attached. HbA1c is measured by a blood test that is recommended every three months by the American Diabetes Association. HbA1c is an important risk factor for patients with diabetes because of the potential for high blood sugar to lead to complications including kidney failure, blindness, and limb amputation. Complete details related to this example can be found in Zhang et al. [82].

Five classes of glucose-lowering medications that are commonly used to control HbA1c were considered: metformin, sulfonylurea, dipeptidyl peptidase 4 (DPP-IV) inhibitors, glucagon-like peptide-1 (GLP-1) agonists, and insulin. Insulin is normally the last line of treatment because of the quality of life impact of having daily injections. In our model, we assumed that once insulin was initiated, HbA1c was controlled at a physician-recommended level of 7%. We also assumed that medications other than insulin had an additive effect in reducing HbA1c.

To estimate the three-month HbA1c transition probabilities, we used anonymized laboratory and pharmacy claims data as described in Zhang et al. [82]. This data set included longitudinal data for HbA1c tests over a multiyear time horizon and pharmacy claims data that provided the frequency and amount of medication refills for all diabetes medications. We identified $37,501$ eligible patients meeting standard criteria for having type 2 diabetes. For these patients, we selected all pairs of records such that the period between tests was between 2.5 and 3.5 months, and the patient was not on insulin during that period. This selection

resulted in 30,249 pairs (multiple pairs permitted per patient). We assumed transition probabilities are stationary over time.

For each medication, we selected patients who had at least one HbA1c measurement within three months before and after initiation of the medication, and who were treated with this medication for at least three consecutive months. For each medication, $m = 1, \ldots, 5$, we calculated the pretreatment HbA1c and the posttreatment HbA1c for all selected patients. We used the mean change in HbA1c observations during the three-month intervals before and after initiation of medication m to estimate the treatment effect in terms of proportional change in HbA1c, denoted by $\omega(m)$. Next, we used the treatment effect and the observed HbA1c value, denoted by $A1c_i^t$ for patient i at epoch t, to estimate the natural HbA1c values in the absence of medication, which we denote by $\bar{A1}c_i^t$:

$$\bar{A1}c_i^t = \frac{A1c_i^t}{1-\omega(m)}, \forall i, t.$$

We subsequently discretized the continuous natural $\bar{A1}c$ values into 10 discrete states using deciles of the empirical distribution to define the interval for each discrete state. For each interval defined by the deciles, we computed the conditional mean and used it as the point estimate of HbA1c for the state associated with the interval. For any two consecutive states, s and s', we denote the total number of transitions from state s to state s' as $n_{s,s'}$. The maximum likelihood estimate of the transition probability is estimated as follows:

$$p(s'|s) = \frac{n_{s,s'}}{\sum_{s' \in \mathcal{S}} n_{s,s'}}, \forall s \in \mathcal{S} \backslash \{\mathcal{D}\},$$

where S is the set of HbA1c states in this example.

The above estimation procedure provides statistical estimates of the transition probabilities among transient health states as described in Section 3. The transitions from health states to the absorbing (disease complications including kidney failure, blindness, and amputation) state were estimated using the United Kingdom Prospective Diabetes Study (UKPDS) outcomes model (Stevens [71]). The UKPDS model is a well-known *survival model* that estimates the probability of future complications based on established risk factors that include age, gender, ethnicity, body mass index, blood pressure, cholesterol, and HbA1c. For this study, which focused on HbA1c control, all risk factors except HbA1c were assumed to be constant over time. Furthermore, the probability of death from other causes was estimated based on the Centers for Disease Control and Prevention (CDC) mortality tables (Anderson and Smith [12]).

The reward function, $r_t(s_t, h_t, a_t)$, was defined as follows:

$$r_t(s_t, h_t, a_t) = \begin{cases} 0.25\big(1 - D^{\text{hyper}}(s_t, a_t)\big)\big(1 - D^{\text{med}}(s_t, h_t, a_t)\big), & \forall s_t \in \mathcal{S} \backslash \{\mathcal{D}\}, h_t \in \mathcal{H}_t, a_t \in \mathcal{A}; \\ 0, & \text{otherwise}, \end{cases} \tag{8}$$

where 0.25 is the three-month length of the time interval $(t, t+1]$ expressed in years, $D^{\text{hyper}}(s_t, a_t)$ is the disutility of daily hyperglycemia symptoms associate with high blood sugar (e.g., headaches, fatigue, frequent urination) when the patient is in state s_t and takes action a_t during epoch $(t, t+1]$, and $D^{\text{med}}(s_t, h_t, a_t)$ is the disutility of taking medications over time interval $(t, t+1]$ that were initiated at or before epoch t. If the patient is on more than one medication, $D^{\text{med}}(s_t, h_t, a_t)$ is the sum of individual medication disutilities corresponding to medication history up to epoch t, h_t, and the most recent action a_t.

The initial HbA1c state distributions at the first decision epoch, mean HbA1c values at diagnosis, and HbA1c state transition probability matrices for men and women are shown in Tables 1 and 2. As mentioned earlier, we considered three-month decision epochs. The time horizon was assumed to begin at the median age of diagnosis of type 2 diabetes (55.2 for women, 53.6 for men; CDC [13]) and ended at age 100 (for patients who survive to the end of

Table 1. Glycosylated hemoglobin (HbA1c) used in the MDP model for women. The table includes the HbA1c range definition at diagnosis, the mean natural HbA1c values for each HbA1c state at diagnosis (before initiating medication), the initial HbA1c distributions at diagnosis, and three-month HbA1c transition probability matrices (TPMs) 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	0.6379	0.3042	0.0481	0.0088	0.0010	0	0	0	0	0
HbA1c state 2	0.1717	0.5085	0.2692	0.0412	0.0064	0.0020	0	0	0	0.0010
HbA1c state 3	0.0299	0.1731	0.5213	0.2258	0.0374	0.0085	0.0018	0.0004	0.0011	0.0007
HbA1c state 4	0.0114	0.0538	0.2830	0.4167	0.1716	0.0446	0.0114	0.0029	0.0021	0.0025
HbA1c state 5	0.0048	0.0240	0.1055	0.2740	0.3329	0.1678	0.0568	0.0199	0.0055	0.0089
HbA1c state 6	0.0045	0.0116	0.0491	0.1438	0.2482	0.2768	0.1598	0.0661	0.0268	0.0134
HbA1c state 7	0.0015	0.0120	0.0316	0.0648	0.1687	0.2364	0.2184	0.1370	0.0768	0.0527
HbA1c state 8	0.0043	0.0065	0.0281	0.0562	0.0864	0.1533	0.1879	0.1965	0.1555	0.1253
HbA1c state 9	0	0.0166	0.0194	0.0332	0.0831	0.1357	0.1662	0.1717	0.1828	0.1911
HbA1c state 10	0.0078	0.0111	0.0277	0.0532	0.0831	0.0920	0.0854	0.0976	0.1042	0.4379

the horizon), after which we assumed the same course of treatment for the remainder of the patient's life. We chose age 100 for two reasons: first, because the average life expectancy after 100 years old is only 2.24 years for women and 2.05 years for men, as reported in Anderson and Smith [12], and second, because the probability of having no macro- or microvascular event or death occur until age 100 is very low. The discount factor λ was chosen to be $\lambda = 1$ to avoid discounting life years. A complete list of the remaining model input sources can be found in Table 3.

Table 2. Glycosylated hemoglobin (HbA1c) used in the MDP model for men. The table includes the HbA1c range definition at diagnosis, the mean natural HbA1c values for each HbA1c state at diagnosis (before initiating medication), the initial HbA1c distributions at diagnosis, and three-month HbA1c transition probability matrices (TPMs) for men.

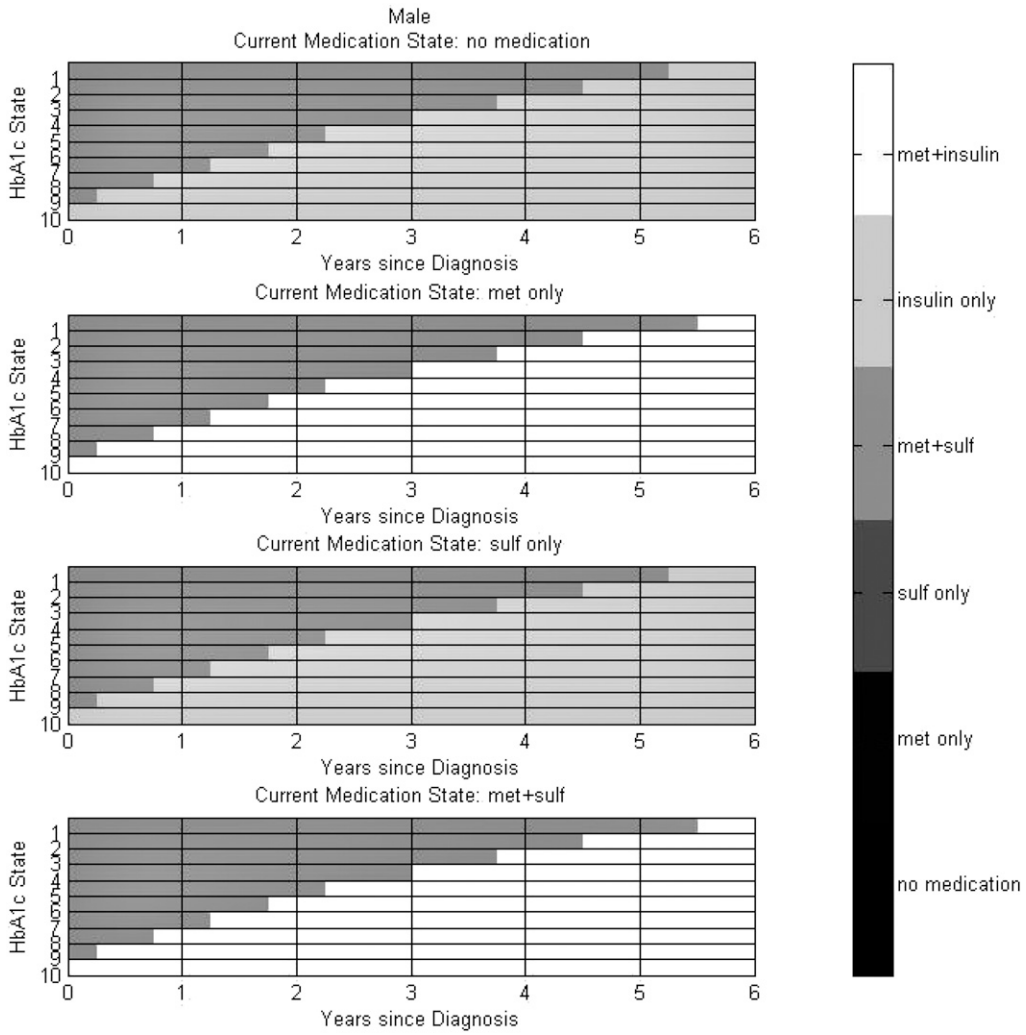
	HbA1c state									
	1	2	3	4	5	6	7	8	9	10
HbA1c range	<6	[6,6.5)	[6.5,7)	[7,7.5)	[7.5,8)	[8,8.5)	[8.5,9)	[9,9.5)	[9.5,10)	≥ 10
Mean HbA1c value (%)	5.69	6.25	6.73	7.24	7.74	8.24	8.74	9.21	9.73	11.59
Initial HbA1c distribution	0.0694	0.1388	0.1968	0.1626	0.1138	0.0919	0.0619	0.0424	0.0328	0.0896
TPM										
HbA1c state 1	0.6245	0.2885	0.0685	0.0093	0.0034	0.0025	0.0008	0.0008	0	0.0017
HbA1c state 2	0.1574	0.4949	0.2953	0.0402	0.0072	0.0038	0.0004	0	0.0004	0.0004
HbA1c state 3	0.0349	0.2061	0.4715	0.2279	0.0441	0.0078	0.0024	0.0012	0.0024	0.0018
HbA1c state 4	0.0130	0.0592	0.2462	0.4014	0.1971	0.0549	0.0166	0.0043	0.0029	0.0043
HbA1c state 5	0.0098	0.0237	0.1058	0.2606	0.3029	0.1852	0.0686	0.0243	0.0083	0.0108
HbA1c state 6	0.0058	0.0134	0.0645	0.1335	0.2313	0.2888	0.1514	0.0550	0.0294	0.0268
HbA1c state 7	0.0104	0.0142	0.0455	0.0796	0.1308	0.2284	0.2351	0.1422	0.0645	0.0493
HbA1c state 8	0.0111	0.0249	0.0456	0.0526	0.0982	0.1674	0.1840	0.1646	0.1328	0.1189
HbA1c state 9	0.0125	0.0233	0.0412	0.0376	0.0789	0.1057	0.1595	0.1792	0.1344	0.2276
HbA1c state 10	0.0098	0.0249	0.0537	0.0688	0.0629	0.0799	0.0911	0.0996	0.1134	0.3958

Table 3. Sources of inputs for the type 2 diabetes model.

Model input	Source
Probabilities among HbA1c states	Claims data set with linked laboratory data (Zhang et al. [82])
Probabilities of adverse events	UKPDS outcome model (Clarke et al. [20])
Probability of death from other causes	CDC mortality tables (Anderson and Smith [12])
End-of-horizon reward	CDC life expectancy tables (Anderson and Smith [12])
Utility of medications	Sinha et al. [67]

Consistent with clinical practice, we assumed the optimal policy for patients who are on insulin is to continue using insulin for their remaining lifetime. Figure 4 shows the optimal policies for patients who are not on insulin, including patients not on any medications, patients on metformin only, patients on sulfonylurea only, and patients on metformin and sulfonylurea

Figure 4. The first six years of the optimal policy from diagnosis of diabetes for men who are not on insulin, including patients not on any medications, patients on metformin only, and patients on metformin and sulfonylurea.



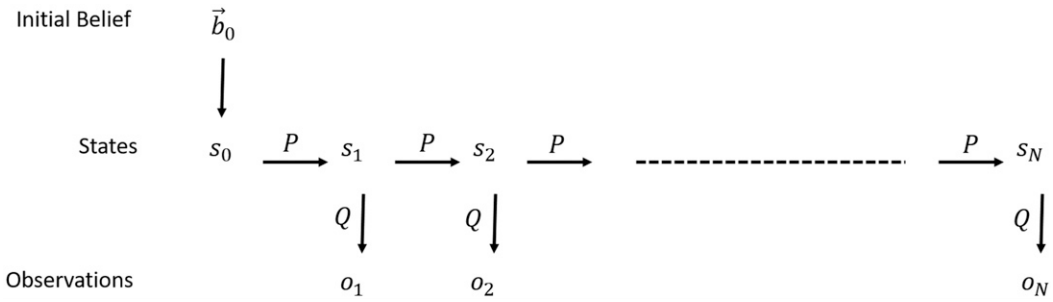
together. The other medications considered, DPP-IV inhibitors and GLP-1 agonists, were not part of the optimal policies. We found that the optimal policies are of control-limit type although the HbA1c transition probability matrices do not satisfy the IFR property exactly. The optimal sequence to initiate medications is the same for men and women, but the time to initiate each medication is different. At the time of diagnosis when patients are not on any medication, the optimal action for those patients with HbA1c less than 10% is to initiate metformin and sulfonylurea together; for those patients with HbA1c $\geq$10%, the optimal action is to initiate insulin immediately. All patients eventually start insulin as a result of the deterioration of glycemic control over time, as suggested by the IFR property being nearly satisfied.

4.2. Model Parameterization for POMDPs

Diseases with latent stages have health states that are not directly observable and are best represented by a *hidden Markov model*. The term "hidden" refers to the fact that the exact health state of the patient is unknown, but observations provide information about the belief the patient is in a given (hidden) core state. Thus, a hidden Markov model is a POMDP without actions or rewards, so developing a hidden Markov model is an important step in formulating a POMDP. To develop a hidden Markov model, it is necessary to estimate the model parameters from observable covariates. This estimation is done using longitudinal data for a population that has received screening tests or diagnostic tests that provide observations that are informative about how likely the patient is to be in a certain health state at multiple time points. In most contexts there is some guideline specifying a recommended starting age, stopping age, and frequency of screening tests. However, it is often the case that observed data deviate from recommended guidelines. This issue can be viewed as a missing data problem and can be addressed using the expectation-maximization (EM) algorithm (Dempster et al. [24]). The EM algorithm iteratively generates model estimates that in theory converge to a maximum likelihood estimate of the hidden model parameters, which include the core state transition probabilities P, information matrix Q, and initial belief b_0, as illustrated in Figure 5. In practice, missing data may be a source of bias because missingness in screening data can be informative (e.g., such as when "sicker" patients receive more frequent screening). In some cases, missing data points or time intervals between data points are informative and may be included as observation variables in the model. We illustrate the major steps of the model formulation with the following example.

4.2.1. Example: Active Surveillance of Low-Risk Prostate Cancer. Active surveillance is commonly recommended for patients with low-risk prostate cancer, as defined by tumor pathology using the Gleason score, a discrete score assigned by a pathologist that differentiates prostate cancer based on the risk of metastases. Active surveillance involves

Figure 5. Illustration of the state transition and observation process for a patient with N observations from the perspective of estimating the parameters for a hidden Markov model including the initial (prior) belief $\vec{b}_0$, state transition probability matrix P, and the information matrix Q.



routine biopsies for patients to confirm whether their cancer continues to be low risk, or whether it has *progressed* to high-risk cancer that should be treated. However, biopsies involve sampling of the prostate using (typically 12) hollow-core needles, and therefore biopsies may fail to identify the presence of a high-grade tumor as a result of sampling error. In this section, we provide an example of a POMDP model for finding an optimal policy for when to refer patients for biopsy. We parameterized the POMDP model in two stages. In the first stage, we estimated the parameters of the hidden Markov model for the unobservable core states of the POMDP. These parameters were computed using the Baum–Welch algorithm, a special case of the EM algorithm, which we describe below. In the second model parameterization stage, transition probabilities from the core states to observable states were estimated using data based on a review of the literature on prostate cancer. The observable states represent treatment, progression to metastatic cancer, and death from any cause. This second stage was necessary because of the low rate of observations of these major events over the limited (10-year) time frame of the longitudinal data set used to estimate the hidden Markov model.

In the context of active surveillance for prostate cancer, a hidden Markov model has core health states defined by prognostic cancer grade groups based on Gleason score. The term "hidden" refers to the fact that the exact health state of the patient is unknown in the absence of surgical removal of the prostate, known as *prostatectomy*. Transition probabilities determine the probability of progression from a low to a high-grade cancer state, which, if detected, is treated by surgery or radiation therapy. We based the model on one-year time periods between state transitions, to be consistent with the highest proposed frequency of biopsies in the literature, and because that was the planned frequency of biopsies in the Johns Hopkins study. The data set was made up of 1,499 patients who initiated active surveillance. The cohort of patients was followed over a 10-year period.

We indexed annual decision epochs as $t = 0, 1, \ldots, T-1$, where $t = 0$ denotes the initial year of diagnosis of a patient with low-risk prostate cancer, before the start of the decision process, which begins at epoch $t = 1$. The model state at epoch t is denoted by $s_t \in \mathcal{S} \equiv \{S_L, S_H\}$, where S_L denotes patients with low-grade cancer and S_H denotes patients with high-grade cancer. Because patients in the high-grade state do not return to the low-grade state, the transition probability matrix is that of an absorbing Markov chain:

$$P = \begin{bmatrix} p(S_L|S_L) & p(S_H|S_L) \\ 0 & 1 \end{bmatrix}.$$

At $t = 0$, patients begin active surveillance under the assumption that they are in state S_L; however, because of a biopsy sampling error, they could be in state S_H. We let $\vec{b}_0 = (b_{0,S_L}, b_{0,S_H})$ denote the initial belief vector of patients in states S_L and S_H at their first surveillance biopsy. The model has observation $o_t \in \mathbb{O} \equiv \{O_-, O_+\}$ at epoch t, where O_- denotes a biopsy observation that indicates low-risk cancer and O_+ denotes a biopsy observation that indicates high-risk cancer. However, biopsies are imperfect as a result of sampling error, and the following matrix denotes the conditional probability of biopsy observations O_- and O_+:

$$Q = \begin{bmatrix} q(O_-|S_L) & q(O_+|S_L) \\ q(O_-|S_H) & q(O_+|S_H) \end{bmatrix}.$$

If a biopsy result is O_+, the patient exits the system and receives treatment according to standard clinical protocols. Collectively, we denote the model parameters for the hidden Markov model by $\mathcal{L} \equiv (\vec{b}_0, P, Q)$. Figure 5 illustrates the stochastic active surveillance process.

Algorithm 3 (Baum–Welch Algorithm for Hidden Markov Model Parameter Estimation)
1: **Input:** Initiate model parameter estimates $\mathcal{L}^0$.
2: Compute $\mathbb{P}(O|\mathcal{L}^0)$ using Equation (9).
3: Compute $\mathcal{L}^1$ using the update equations in (10).
4: Compute $\mathbb{P}(O|\mathcal{L}^1)$ using Equation (9).

5: $k \leftarrow 1$
6: **while** $\mathbb{P}(O|\mathscr{L}^k) - \mathbb{P}(O|\mathscr{L}^{k-1}) > Tolerance$ **do**
7: $v \leftarrow v+1$
8: Compute $\mathscr{L}^k$ using the update equations in (10).
9: Compute $\mathbb{P}(O|\mathscr{L}^k)$.
10: **end while**

Maximum likelihood estimates of $\mathscr{L}$ were obtained using the Baum–Welch algorithm (Algorithm 3). The Baum–Welch algorithm is an iterative algorithm that combines forward and backward passes on a longitudinal observation sequence to find the choice of $\mathscr{L}$ that maximizes the likelihood of observing the collection of sequences. In our application, we have biopsy results for $v = 1, \ldots, N$ patients, where $N = 1,499$. Each patient v has an observation sequence, $O^{(v)} = \{o_1^{(v)}, o_2^{(v)}, \ldots, o_{T_v}^{(v)}\}$, which represents a patient's biopsy results over T_v time periods. We denote the set of N observation sequences as $O \equiv \{O^{(1)}, O^{(2)}, \ldots, O^{(N)}\}$. Thus, our goal is to find the model $\mathscr{L}$ that maximizes

$$\mathbb{P}(O \,|\, \mathscr{L})) = \prod_{v=1}^{N} \mathbb{P}\big(O^{(v)} | \, \mathscr{L}\big), \tag{9}$$

where we assume that observation sequences among patients are independent. To describe the steps of the Baum–Welch algorithm, we denote elements of matrices P and Q as $p(j|i)$ and $q(o|i)$, respectively, dropping the subscript t because we assume the matrices are stationary. We begin by defining the *forward variable*, $\alpha_t^{(v)}(i)$, of the Baum–Welch algorithm as

$$\alpha_t^{(v)}(i) = \mathbb{P}(o_1^{(v)}, o_2^{(v)}, \ldots, o_t^{(v)}, s_t = S_i|\mathscr{L}), i = 1, 2,$$

where $S_1 \equiv S_L$ and $S_2 \equiv S_H$, and $\alpha_t^{(v)}(i)$ is the probability of observing the partial observation sequence until time t and being in state S_i at time t, given the model $\mathscr{L}$. Forward recursion is used to efficiently solve for $\alpha_t^{(v)}(i)$:

$$\alpha_1^{(v}(i) = b_{0,S_i} q\big(o_1^{(v)}|i\big), \quad i = 1, 2, \nu = 1, \ldots, N,$$

$$\alpha_{t+1}^{(v)}(j) = \left(\sum_{i=1}^{2} \alpha_t^{(v)}(i) p(j|i)\right) q\big(o_{t+1}^{(v)}|j\big), \quad 2 \le t \le T_v - 1, \, j = 1, 2, \, \nu = 1, \ldots, N.$$

Next, we define the *backward variable*, $\beta_t^{(v)}(i)$, as follows:

$$\beta_t^{(v)}(i) = \mathbb{P}\Big(o_{t+1}^{(v)}, o_{t+2}^{(v)}, \ldots, o_{T_v}^{(v)}|\mathscr{L}, s_t = S_i\Big),$$

which is the probability of the partial observation sequence from $t+1$ to T_v, given the model $\mathscr{L}$, and given that patient v is in state S_i at time t. Backward recursion can be used to efficiently solve for $\beta_t^{(v)}(i)$ given that $\beta_{T_v}^{(v)}(i) = 1$, and using the following recursive equations:

$$\beta_t^{(v)}(i) = \sum_{j=1}^{2} p(j|i) q\big(o_{t+1}^{(v)}|j\big) \beta_{t+1}^{(v)}(j), \quad t = T_v - 1, \ldots, 1, \; i = 1, 2. \, \nu = 1, \ldots, N.$$

To define the iterative procedure that underlies the Baum–Welch algorithm, we let $\xi_t^{(v)}(i, j)$ denote the probability of patient v being in state S_i at epoch t, and then state S_j at epoch $t+1$, given the model $\mathscr{L}$ and the observation sequence $O^{(v)}$:

$$\xi_t^{(v)}(i,j) = \mathbb{P}\big(s_t = S_i, s_{t+1} = S_j|\mathscr{L}, O^{(v)}\big) = \frac{\alpha_t^{(v)}(i) p(j|i) q\big(o_{t+1}^{(v)}|j\big) \beta_{t+1}(j)}{\mathbb{P}(O^{(v)}|\mathscr{L})},$$
$$t = T_v - 1, \ldots, 1, \; i = 1, 2.$$

The probability of patient v being in state S_i at time t, given the model $\mathscr{L}$ and the observation sequence $O^{(v)}$, can written as follows:

$$\gamma_t^{(v)}(i) = \sum_{j=1}^{2} \xi_t^{(v)}(i,j), \quad i = 1, 2,\ \nu = 1, \ldots, N.$$

From the above definitions, we can write the following update formulas for the model parameters, which iteratively improve $\mathbb{P}(O \mid \mathscr{L})$, as using the above forward and backward equations, as outlined in Algorithm 3:

$$\bar{p}(j|i) = \frac{\sum_{v=1}^{N} \frac{1}{P_v} \sum_{t=1}^{T_v - 1} \xi_t^{(v)}(i,j)}{\sum_{v=1}^{N} \frac{1}{P_v} \sum_{t=1}^{T_v - 1} \gamma_t^{(v)}(i)}, \ \bar{q}(l|j) = \frac{\sum_{v=1}^{N} \frac{1}{P_v} \sum_{\{\forall t | O_t^{(v)} = O_-\}}^{T_v - 1} \gamma_t^{(v)}(j)}{\sum_{v=1}^{N} \frac{1}{P_v} \sum_{t=1}^{T_v - 1} \gamma_t^{(v)}(j)}, \ \bar{b}_{0,i} = \sum_{v=1}^{N} \frac{\gamma_1^{(v)}(i)}{P_v}, \tag{10}$$

where $P_v \equiv \mathbb{P}(O^{(v)}|\mathscr{L})$ is used for conciseness. The update equation for $\bar{p}(j|i)$ calculates the expected ratio of the number of transitions from state S_i to state S_j divided by the expected number of transitions from state S_i. The update equation for $\bar{q}(l|j)$ calculates the expected number of times a patient is in state S_j and observes l divided by the expected number of times a patient is in state S_j. Finally, the update equation for $\bar{b}_{0,i}$ is the expected proportion of times a patient is in state S_i at $t = 0$.

The Baum–Welch algorithm (Algorithm 3) uses the above formulas to update the parameters of $\mathscr{L}$ iteratively. A proof of convergence for the Baum–Welch algorithm follows from the convergence guarantees for the general EM algorithm; however, as with the EM algorithm, convergence to a local optimum is possible because the maximization problem is not strictly convex, and thus the limiting point for the sequential updates may be sensitive to the starting point. For this reason, we conducted a sensitivity analysis using various starting points to initiate the algorithm to confirm the robustness of the final solution. We also analyzed data generated by sampling from known models with selected data elements missing at random to confirm convergence of the Baum–Welch algorithm under these conditions.

Applying Algorithm 3 to the 1,499 patients who initiated active surveillance in the Johns Hopkins study, with a tolerance of 10^{-6} for changes in the likelihood function from one iteration to the next, we estimated the annual progression rate from low to high risk to be $p(S_H|S_L) = 0.04$. The sensitivity and specificity of biopsy for high-risk cancer were estimated to be $q(O_+|S_H) = 0.61$ and $q(O_-|S_L) = 0.986$, respectively. Thus, biopsy identified patients with high-risk disease approximately 61% of the time. For patients with low-risk disease, biopsies correctly find no evidence of high-risk disease 99% of the time. Finally, the initial proportion of patients misclassified at the time of diagnosis, as a result of inaccuracy in the biopsy sampling process, was estimated to be $\vec{b} = (0.902, 0.098)$. Therefore, about 10% of patients who start active surveillance in this cohort have an undiagnosed high-risk disease that warrants treatment.

In the second stage of model parameterization, we added three additional observable states: treatment (S_T), metastatic cancer (S_M), and death from any cause (S_D). The belief vector for the POMDP for decision epoch t is $\vec{b}_t = (b_{t,S_L}, b_{t,S_H}, b_{t,S_T}, b_{t,S_M}, b_{t,S_D})$. The additional states were added based on post hoc analysis using data from the literature because these endpoints were not available in the active surveillance study data as a result of limited follow-up time. The relevant parameters used to estimate the transition probabilities are as follows:

δ_t: annual other-cause mortality rate, $S_L/S_H/S_T/S_M \rightarrow S_D$.
γ_t: annual metastasis rate with treatment, $S_T \rightarrow S_M$.
$\bar{\gamma}_t$: annual metastasis rate without treatment, $S_H \rightarrow S_M$.
ϕ_t: annual prostate death rate, $S_M \rightarrow S_D$.
θ_t: annual progression rate, $S_L \rightarrow S_H$.

The reward function is based on QALYs where the annual utility for a year of life in unobservable states (S_L, S_H) is 1 and the annual utility for state S_D is 0. The remaining parameters are the following:

u_T: annual utility for life after treatment (S_T).
u_M: annual utility for life with metastatic cancer (S_M).

The reward function also includes additional disutility in the year of treatment, κ, and disutility for biopsy, μ, in the year of the procedure. The time horizon for the POMDP assumed diagnosis at age 50 ($t = 1$) and the last period was age 100 ($T = 51$) with a terminal reward equal to expected life span for patients alive in the last epoch. When a patient enters one of the completely observable states S_T, S_M, S_D, there are no remaining actions, and the rewards are computed via a Markov reward process with the following rewards:

$$\begin{aligned}
&\bar{R}_t(S_T) = u_T + \delta_t \bar{R}_{t+1}(S_D) + (1-\delta_t)\gamma_t \bar{R}_{t+1}(S_M) + (1-\delta_t)(1-\gamma_t)\bar{R}_{t+1}(S_T),\\
&\bar{R}_t(S_M) = u_M + (\delta_t + (1-\delta_t)\phi_t)\bar{R}_{n+1}(S_D) + (1-\phi_t)(1-\delta_t)\bar{R}_{t+1}(S_M),\\
&\bar{R}_t(S_D) = 0.
\end{aligned}$$

The boundary condition for the observable states is

$$\bar{R}_T(S_T) = u_T \ell_T, \bar{R}_T(S_M) = u_M \ell_M, \bar{R}_T(S_D) = 0,$$

where ℓ_T and ℓ_M are expected life spans for patients who are alive in the last period T. All parameter values associated with the second stage of parameterization of the POMDP model and their sources are provided in Table 4.

The optimality equations for the POMDP model, which maximizes total expected QALYs, can be expressed as follows:

$$v_t(\vec{b}_t) = \max \begin{cases} r_t(\vec{b}_t, W) + \lambda\big(\delta_t \bar{R}_{t+1}(S_D) + (1-\delta_t)\bar{\gamma}\hat{b}_t^{\emptyset}(S_H)\bar{R}_{t+1}(S_M) & \\ +(1-\delta_t)\big(1-\bar{\gamma}_t\big)\hat{b}_t^{\emptyset}(S_H) v_{t+1}(b_{t+1}^{\emptyset})\big), & a_t = W \\ r_t\big(\vec{b}_t, B\big) + \lambda\big(\mathbb{P}\big(O_+|\vec{b}_t, B\big)\big(\delta_t \bar{R}_{t+1}(S_D) + (1-\delta_t)\gamma_t \hat{b}_t^{+}(S_H)\bar{R}_{t+1}(S_M) & \\ +(1-\delta_t)(1-\gamma_t)\bar{R}_{t+1}(S_T) - \kappa) + & \\ \mathbb{P}\big(O_-|\vec{b}_t, B\big)\big(\delta_t \bar{R}_{t+1}(S_D) + (1-\delta_t)\bar{\gamma}_t \hat{b}_t^{-}(P)\bar{R}_{t+1}(S_M) & \\ +(1-\delta_t)\big(1-\bar{\gamma}_t \hat{b}_t^{-}(S_H)\big) v_{t+1}\big(b_{t+1}^{-}\big)\big), & a_t = B, \end{cases}$$

where $\bar{R}_T(\vec{b}_T) = 3.34$, which is the expected life span of an individual at age 100 according to U.S. life tables, $r_t(\vec{b}_t, W) = 1$, and $r_t(\vec{b}_t, B) = 1 - \mu$. The vectors $\hat{b}_t^{\emptyset}$, $\hat{b}_t^{-}$, $\hat{b}_t^{+}$ are belief vectors immediately after the observation $\emptyset, -, +$, respectively, where $\emptyset$ indicates no biopsy was performed. We denote by $\mathbb{P}(o_t|\vec{b}_t, B)$ the probability of observation o_t in period t given belief vector $\vec{b}_t$ and action B. Specifically,

$$\hat{b}_{t,S_H}^{\emptyset} = b_{t,S_H},\ \hat{b}_{t,S_H}^{+} = 1,\ \hat{b}_{t,S_H}^{-} = \frac{q(O_-|S_H) b_{t,S_H}}{q(O_-|S_H) b_{t,S_H} + b_{t,S_L}}.$$

Figure 6 shows the set of α-vectors obtained from solving the POMDP using Algorithm 2 for the optimal value function at $t = 0$, which represented expected quality-adjusted life span for a man diagnosed with low-risk prostate cancer at age 50. Each α-vector is associated with one of the two actions: biopsy or defer biopsy. The x axis represents the element of the belief vector, b_{t,S_H}, corresponding to the patient being in the high-risk cancer state. At each point on the x axis, there is an α-vector that corresponds to the optimal action for that belief point. Thus, Figure 6 provides a complete representation of the optimal policy for the POMDP.

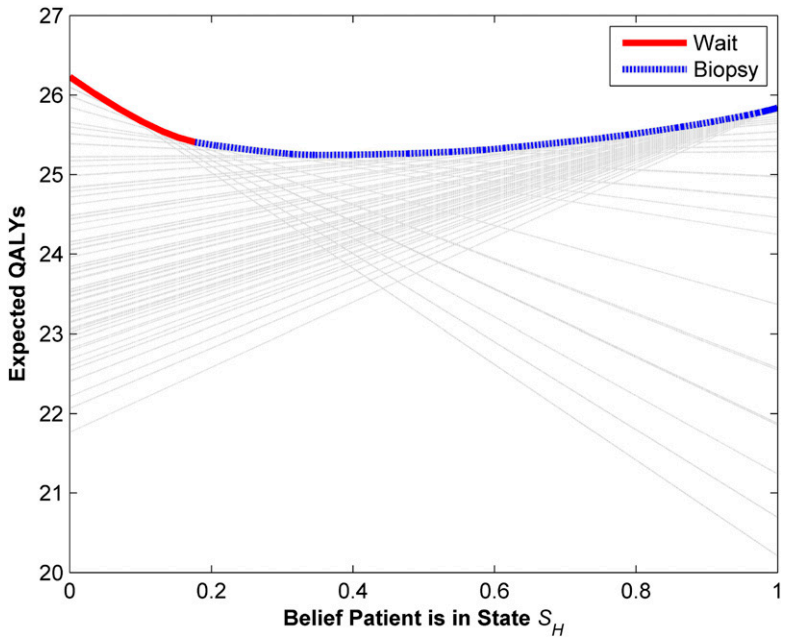
Table 4. Summary of parameters for the POMDP model.

Parameter	Meaning	Source	Value
δ_t	Annual mortality rate of other-cause diseases	Anderson and Smith [12]	0.002∼0.299 (Age specific)
$\bar{\gamma}_t$	Annual metastasis rate of untreated PCa	Ghani et al. [30]	0.069
γ_t	Annual metastasis rate of treated PCa	Zhang et al. [81]	0.006
ϕ_t	Annual PCa death rate of metastasized PCa	Anderson and Smith [12]	0.07∼0.074 (Age specific)
$1-\mu_T$	Annual disutility for posttreatment	Heijnsdijk et al. [38]	0.05
$1-\mu_M$	Annual disutility for metastasis	Heijnsdijk et al. [38]	0.4
μ	Instantaneous disutility for biopsy	Chhatwal et al. [17], Kulkarni et al. [42]	0.05
κ	Instantaneous disutility for treatment	Heijnsdijk et al. [38]	0.247

5. Data Sources and Model Uncertainty

Parameterization of models, such as the MDP and POMDP models of the previous section, relies on data. There are many sources of data on chronic diseases such as research study data, insurance claims data (private and public), hospital and outpatient clinic data, vital statistics collected by government agencies (e.g., Centers for Disease Control and Prevention in the United States), and many others. The U.S. National Institutes of Health (NIH) is the principal agency in the United States that provides research funding for studies of a vast range of diseases. Many of these studies share their data with other researchers according to the *data sharing plan* that is required as part of the initial grant application. The following NIH policy guidance statement explains what researchers can expect [55]: "In NIH's view, all data should be considered for data sharing. *Data should be made as widely and freely available as possible*

Figure 6. (Color online) Illustration of all of the nondominated α-vectors for the optimal value function of the POMDP for active surveillance at epoch $t = 1$.



while safeguarding the privacy of participants, and protecting confidential and proprietary data" (emphasis in original). Not all data are available upon request for reasons such as a special need for privacy, or in situations where data are proprietary; however, many data sets are available either through a defined process for data sharing or a formal data use agreement with a project investigator. Moreover, there are large observational data sets that are available at a cost from public and private initiatives in the United States that aggregate data on large portions of the population (e.g., Medicare data).

Observational data can be challenging to work with, as opposed to prospectively collected research study data, because there is no control over the health interventions patients receive in such cases, making it difficult to estimate a natural history model for a given disease. For problems with a single one-time intervention (i.e., stopping-time problems), patients transition to an absorbing state representing posttreatment survival; thus, it may not be necessary to consider the effect of treatment on the transition probabilities in such cases. When there are multiple interventions (e.g., multiple medications, as in the diabetes example of Section 4), the influence of interventions on transition probabilities becomes important. The previous examples in the context of type 2 diabetes and prostate cancer provide examples of ways to estimate natural history models.

Often, the effect of interventions is to lower the probability of having an adverse event associated with the disease. For many common diseases, statistical survival models exist that can be used to estimate the probability of an adverse event. For instance, statistical models for type 2 diabetes include the Framingham model (Anderson [6]), the UKPDS model (Stevens et al. [71]), and the American College of Cardiology/American Heart Association pooled risk calculator (Goff et al. [32]). These models predict the probability of diabetes complications including cardiovascular events (stroke and coronary heart disease), kidney failure, and blindness. Model inputs include gender, race, family history, and metabolic factors such as cholesterol, blood pressure, and blood glucose. Health interventions modify the inputs to these models and thus reduce the probability of transitioning to a complication state. When using survival models in this way, it should be acknowledged that there is an implied assumption of a causal relationship between the change in a risk factor (e.g., cholesterol, blood pressure) and the risk of complications (e.g., heart attack, stroke).

5.1. Limitations of Model Estimation Because of Limited or Biased Data

The example of Section 4.1.1 describes a simple approach for estimating transition probabilities that may be appropriate in situations in which there are enough data to identify a suitably large subset of patients that initiated each of the interventions under study and for which observations of the risk factor in question occur at high frequency. For data sets with less frequent observations, imputation methods can provide a means to approximate individual patient's health status over time. For example, Shechter et al. [65] use curve fitting (smoothing splines) to impute CD4 count between laboratory measures made at different points in time. This imputation procedure is one of many approaches to addressing the commonly encountered challenge of missing data.

The simplest approach to dealing with missing data is to discard data for any patient that has missing data, which is known as a *complete case* analysis. This approach may be appropriate if there are a large number of samples and if the data are believed to be missing *completely at random* (i.e., not associated with some measured or unmeasured covariate). A straightforward approach that avoids deleting patient records is *single imputation*, which replaces the missing value with a plausible estimate, such as the population mean (e.g., a patient with missing heart rate would have his heart rate replaced with the mean over the remaining patients). More advanced methods, such as *stochastic regression imputation*, use regression models to impute values by including relevant covariates when fitting the model and then sampling the normally distributed residual term. The text by Enders [26] is an excellent introductory resource for learning about missing data analysis.

Any approach that uses observational data may be biased. For example, treatment estimates can be biased because the population that received treatment is likely to differ from the population that did not. In the above example, this would occur if patients with high HbA1c levels are more likely to receive treatment and if the response of these patients to treatment differs from patients with low HbA1c levels. Regression models can be used to reduce bias by incorporating covariates that have the potential to influence the risk factor under consideration and the decisions to initiate medications that may, in turn, affect the risk factor. Linear regression for the HbA1C example would have HbA1C as the *dependent variable*, and *independent variables* would be age, gender, diabetes medications, time since diagnosis, and any other factors that may explain HbA1c for a particular patient. A multidimensional model such as this is likely to produce more accurate estimates of treatment effects for medications by accounting for the role of covariates that are statistically significant in describing the observations. There are many possible choices of regression models to use, and the best choice depends on the context. Many studies in the context of chronic diseases use longitudinal data for multiple patients, often referred to as *panel data*, that provide repeated measures for each patient in the panel. Given the reasonable expectation of variation between patients and correlation of repeated measures within patients, *random effects* models are commonly employed because they incorporate random intercepts and model coefficients that describe variance and covariance of the observed dependent variable over time. A helpful tutorial on random effects models can be found in Littell et al. [47].

Propensity scoring is another well-known method that is used to address bias in observational data. Propensity scoring is one of many types of *matching methods* that consider the influence of patient attributes on the decision to initiate a health intervention. A propensity score is typically based on a *logistic regression* model to predict whether patients will receive treatment on the basis of their attributes. The score is then used to match patients by the factors that influence treatment, which reduces bias by mitigating the influence that patient attributes have on estimates of the outcome variable. D'Agostino [22] provides an excellent introductory tutorial on propensity scoring for bias reduction in the context of estimating treatment effects.

5.2. Quantifying Model Uncertainty

The examples of Section 4 are based on parameters obtained from point estimates derived from longitudinal data, in the case of transition probabilities, or from survey data in the case of disutilities. All of these parameters are subject to statistical variation, and sensitivity analysis should be employed to assess the influence of model parameter uncertainty on any conclusions that are drawn from the model. Approaches for sensitivity analysis vary depending on whether the analysis considers reward parameters or transition probabilities. In some sense, reward parameter uncertainty is more straightforward because the rewards often vary independently as opposed to transition probabilities that must lie in the unit simplex.

Sensitivity analyses for reward parameters typically vary the parameters that define the rewards within some plausible range such as a statistical confidence region or a range of estimates drawn from multiple sources. Re-solving models for different choices of the reward parameters provides information on the influence of changes on the optimal value function and the optimal policy. It is important to consider both of these types of changes because it may be the case that changes in a specific parameter will affect the optimal value function significantly, whereas the optimal policy may not be sensitive to these changes. Tan and Hartman [73] explore the inverse problem of determining the degree to which reward parameters can vary without the optimal policy changing.

Several authors have studied sensitivity to variation in transition probability matrices. For example, Mannor et al. [50] and Goh et al. [33] analyze the variance of the estimated value function for a stationary infinite-horizon Markov model using the closed-form expression for the value function. Approaches suitable for nonstationary finite-horizon models include

bootstrapping (Craig and Sendi [21]) and a Bayesian approach (Briggs et al. [10]) that uses the source data for the original model estimation. Chen et al. [16] describe a framework for conducting sensitivity analysis on the optimal policy of an MDP using a Bayesian approach to sample the MDP's transition probabilities. Zhang et al. [80] propose an approach for sensitivity analysis of transition probabilities based on *Monte Carlo* sampling under the condition that each transition probability has a defined range (e.g., statistical confidence interval, expert opinion), but the source data are not available. For a given row of the transition probability matrix, $p(\cdot \mid \vec{s}_t)$ at epoch t, the elements of the row must satisfy the following conditions:

$$LB_t(\vec{s}_{t+1}|\vec{s}_t) \le p_t(\vec{s}_{t+1}|\vec{s}_t) \le UB_t(\vec{s}_{t+1}|\vec{s}_t), \quad \sum_{s_{t+1}\in\mathcal{S}} p_t(\vec{s}_{t+1}|\vec{s}_t) = 1, \tag{11}$$

where $LB_t(\vec{s}_{t+1}|\vec{s}_t)$ and $UB_t(\vec{s}_{t+1}|\vec{s}_t)$ are the lower and upper bounds on the transition probabilities, respectively. Zhang et al. [80] propose an implementation of Smith's random-direction algorithm for generating a series of uniformly distributed random points within a bounded, convex region (in this case, the intersection of a hyperrectangle and the unit simplex defined by the polytope in Equation (11)) (Smith [69]).

6. Other Models for Sequential Decision Making

The focus of this tutorial has been on MDPs and POMDPs because these are fundamental models that serve as a foundation for other approaches to sequential decision making. There are many important variations these of models and methods that are relevant to the context of medical decision making. In this section, we provide a few brief examples.

Robust MDPs. Approaches for addressing parameter uncertainty have been studied for many years. The basic idea is to find policies that satisfy some measure of robustness with respect to parameter variation. Satia and Lave [62] were among the first to consider MDPs with uncertain parameters. They consider stationary problems in the context of a stochastic game in which one player decides on the policy and the other selects the model parameters from some defined region. White and El-deib [77] consider uncertainty in reward parameters in the context of infinite-horizon MDPs with the goal of generating the complete set of optimal policies over a convex set of rewards. Later work by the same authors extends the work of Satia and Lave [62] by presenting bounds and approximation methods for the case of an MDP with uncertain transition probabilities in the context of a stochastic game (White and El-deib [78]). More recently, Iyengar [40] and Nilim and Ghaoui [56] have provided an analysis of problems that address the issue of uncertainty in transition probabilities using a min-max approach. They provide an analysis of the problems including differentiating between easy versus hard problems where the dividing line is drawn in part by a commonly employed assumption known as *rectangularity*, which implies independence among rows of the transition probability matrix. Under the rectangularity assumption, the simplest version of a robust MDP is the *interval model* that assumes that independent intervals on the transition probabilities collectively define the *uncertainty set*, $\mathcal{U}(s_t, h_t, a_t)$, within which the transition probability matrix varies. The optimality equations in this context are

$$v_t(s_t, h_t) = \max_{a_t\in\mathcal{A}_t(s_t)} \left\{ r_t(s_t, h_t, a_t) + \lambda \min_{p_t(\cdot|s_t,h_t,a_t)\in\mathcal{U}(s_t,h_t,a_t)} \left\{ \sum_{\forall j\in\mathcal{S}} p_t(j|s_t, h_t, a_t) v_{t+1}(j) \right\} \right\}, \tag{12}$$

for all $s_t \in \mathcal{S}$, $h_t \in \mathcal{H}_t$, and all $t \in \mathcal{T}\backslash\{T\}$. The end-of-horizon boundary condition is $v_T(s_T, h_T) = r_T(s_T, h_T)$ for all $s_T \in \mathcal{S}$ and $h_T \in \mathcal{H}_T$. The minimization problem in Equation (12) is known as the *inner problem*. The computational effort to solve the optimality equations depends in large part on the inner problem, which in turn depends strongly on the structure of the uncertainty set. The rectangularity property guarantees the problem can be decomposed state by state at each decision epoch. However, in some cases, this assumption, which allows the adversary to modify the transition probabilities independently at each decision epoch, may generate policies

that are extremely conservative in their response to uncertainty. An illustration of this can be found in the context of medical treatment decisions in Zhang et al. [80], where the authors show that expected value to go for the true worst-case policy is far more pessimistic than the worst case achieved from Monte Carlo sampling of transition probability matrices. Wiesmann et al. [79] examine the implications of relaxing the rectangularity assumption in various ways and classify the complexity of the resulting problems. Steimle et al. [70] present a new framework that relaxes the rectangularity function and seeks to mitigate the conservative nature of max-min formulations by focusing on an optimal policy that performs well based on a distribution of MDPs.

Reinforcement Learning. Many learning-based methods are focused on sequential decision making under the fundamental assumption that the states and actions of the system are known, but the underlying transition probabilities are not. In this context, the focus becomes one of learning over time through a combination of *exploration* and *exploitation* of the reward response associated with alternative state–action pairs. The learner is assumed to observe the rewards associated with each state–action pair, $r_t(s_t, h_t, a_t)$ at each epoch t. The simplest example of an algorithmic implementation is Monte Carlo policy evaluation. In this setting, state–action pairs are generated using simulation based on a black box (i.e., the underlying model in unknown). The learner does not know the model that generates the state–action pairs, but through a process of sequential experimentation, the ideal policy can be learned. The new policy selected in each iteration is based either on the policy that maximizes expected rewards (exploitation) or on a randomly selected policy (exploration). The latter occurs with some probability ϵ that is selected as an algorithm parameter. The randomized policy $\hat{\vec{\pi}}$ uses the policy obtained from the policy improvement step with probability $1-\epsilon$; otherwise, the policy is determined by randomly selecting actions according to some probability distribution, with probability ϵ. The algorithm converges to the optimal policy asymptotically; however, this is not a practical algorithm because the number of sample paths is infinite. Using a finite number of samples surfaces some important questions about how to trade off statistical error with accuracy in the policy evaluation step. Whereas Algorithm 4 serves as a useful device for explaining the basic concept behind reinforcement learning, many other more efficient approaches have been proposed that take advantage of incremental updating, such as temporal difference learning, Q-learning, and many others. The introductory textbook by Sutton and Barto [72] provides an excellent survey of these methods. Murphy [53] develops a statistical framework for estimating optimal adaptive treatment policies in a sequential decision-making setting. Her work is motivated by the use of data from randomized trials, under the assumption that a stochastic model for changes in a patient's health status over time is unknown, but observational data about the patient's health state and treatment actions are available periodically at discrete time intervals. This work has led to important methodological extensions (see, e.g., Lei et al. [45] and Murphy [54]) and applications in contexts such as depression, drug addiction, hypertension, and warfarin dosing (Anderson et al. [7], Brooner and Kidorf [11], Glasgow et al. [31], Untzer et al. [76]), to name a few examples.

Algorithm 4 (Monte Carlo Policy Iteration)

1: **Input:** Select an initial policy $\vec{\pi}$ for policy evaluation. Choose ϵ for randomization of policy $\hat{\pi}$.
2: **Policy Evaluation:** Randomly select an infinite number of starting pairs (s_0, a_0) and a corresponding sample path of future states and actions. For all $(s_t, h_t, \hat{\pi}(s_t))$ in the sample path compute $v(s_t, h_t, \hat{\pi}(s_t)) = r_t(s_t, h_t, \hat{\pi}(s_t, h_t)) + \lambda \sum_{t'=t}^{T-1} \lambda^t r_{t'}(s_{t'}, h_{t'}, \hat{\pi}(s_{t'}) + r_T(s_T, h_T)$.
3: **Policy Improvement:** For all (s_t, h_t): $\pi(s_t, h_t) = \arg\max\{v(s_t, h_t, \pi(s_t, h_t)\}$.
4: Return to Policy Evaluation.

Simulation-Optimization. Surprisingly, the use of simulation-optimization in medical decision making appears to be quite limited given the degree to which simulation is used for modeling of diseases (Kuntz et al. [43]). Many models in the medical literature are based on Monte Carlo simulation with the goal of evaluating one or more predetermined policies for

a population of patients represented by a *microsimulation model*, so called because such models simulate each patient independently (also called first-order Monte Carlo simulations). Simulation optimization is particularly applicable to cases in which the Markov assumption may not be appropriate and thus may disqualify MDPs as optimization models. One such example is in the optimization of prostate cancer screening policies based on PSA, a biomarker that increases over time following the onset of cancer (Gulati et al. [35]). Underwood et al. [75] use a genetic algorithm with Monte Carlo simulation for evaluation of an iteratively generated *population* of PSA-screening strategies. The authors show that this approach could identify age-dependent screening strategies that outperform "static" policies proposed in the literature. Fu [29] provides an excellent survey of simulation-optimization methods for the interested reader.

7. Conclusions

Modeling approaches for optimizing sequential decision making such as MDPs and POMDPs hold great promise for transforming raw data into optimal policies for health interventions. Early work in the operations research field has focused on model analysis and algorithm development. By contrast, the medical field has focused on the development of stochastic models of diseases using a wide range of data sources. The recent emergence of large observational data sets, funded by public and private initiatives in the United States, combine these different types of data to create longitudinal data on large portions of the U.S. population. However, the full potential of these data are largely untapped because of a lack of data science methods for addressing missing data, data errors, selection bias, verification bias, and other characteristics that confound elicitation of optimal treatment practices from longitudinal observational data. As a result, opportunities abound for research at the intersection of sequential decision making and model estimation that can leverage the discoveries from these two bodies of literature. There is a pressing need to develop approaches for mitigating the influence of uncertainty, bias, and ambiguity that naturally occurs in all mathematical models but that can be particularly cumbersome in a sequential setting because of the resulting loss of convenient structural properties that allow for the "divide-and-conquer" approach commonly employed in dynamic programming. There is also a great need to develop well-validated models that confirm or challenge the strong assumptions that are common in sequential decision making, such as a *risk-neutral* decision maker and the Markov assumption that is commonly employed. Finally, there is a need for success stories in the real-world implementation of sequential optimization approaches in medicine to build trust among decision makers including physicians and patients.

Acknowledgments

This tutorial is also based in part on work supported by the National Science Foundation [Grants CMMI 1462060 and CMMI 1536444]. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. The author is indebted to many collaborators with whom he has worked over the years on topics related to this tutorial. He is also grateful for feedback and help in compiling numerical examples from Weiyu Li, Lauren Steimle, and Zhang Zhang and for feedback from two anonymous reviewers and Douglas Shier, the editor of the 2018 INFORMS TutORials.

References

[1] O. Alagoz, H. R. Hsu, A. J. Schaefer, and M. S. Roberts. Markov decision processes: A tool for sequential decision making under uncertainty. *Medical Decision Making* 30(4):474–483, 2010.

[2] O. Alagoz, L. M. Maillart, A. J. Schaefer, and M. S. Roberts. The optimal timing of living donor liver transplantation. *Management Science* 50(10):1420–1430, 2004.

[3] E. Altman. *Constrained Markov Decision Processes*, Vol. 7. CRC Press, Boca Raton, FL, 1999.

[4] American Diabetes Association. Standards of medical care in diabetes. *Diabetes Care* 41(Suppl. 1): S55–S64, 2018.

[5] R. N. Anderson and B. L. Smith. Deaths: Leading causes for 2002. *National Vital Statistics Reports* 53(17):1–89, 2005.

[6] K. A. Anderson, P. M. Odell, P. W. Wilson, and W. B. Kannel. Cardiovascular disease risk profiles. *American Heart Journal* 121(1, Part 2):293–298, 1991.

[7] J. L. Anderson, B. D Horne, S. M. Steven, A. S. Grove, S. Barton, Z. P. Nicholas, S. F. S. Kahn, et al. Randomized trial of a genotype-guided versus standard warfarin dosing in patients initiating oral anticoagulation. *Circulation* 116(22):2563–2570, 2007.

[8] T. Ayer, O. Alagoz, and N. K. Stout. A POMDP approach to personalize mammography screening decisions. *Operations Research* 60(5):1019–1034, 2012.

[9] D. P. Bertsekas. *Dynamic Programming and Optimal Control*, Vols. 1 and 2. Athena Scientific, Belmont, MA, 1995.

[10] A. H. Briggs, A. E. Ades, and M. J. Price. Probabilistic sensitivity analysis for decision trees with multiple branches: Use of the Dirichlet distribution in a Bayesian framework. *Medical Decision Making* 23(4):341–350, 2003.

[11] R. K. Brooner and M. Kidorf. Using behavioral reinforcement to improve methadone treatment participation. *Science and Practice Perspectives* 1(1):38–47, 2002.

[12] A. Cassandra, M. L. Littman, and N. L. Zhang. Incremental pruning: A simple, fast, exact method for solving partially observable Markov decision processes. D. Geiger and P. P. Shenoy, eds. *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Francisco, 54–61, 1997.

[13] Centers for Disease Control and Prevention. National diabetes fact sheet: National estimates and general information on diabetes and prediabetes in the United States. Technical report, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, Atlanta, 2011.

[14] Centers for Disease Control and Prevention. Chronic disease overview. Accessed April 17, https://www.cdc.gov/chronicdisease/overview/index.htm, 2018.

[15] I. Chadès, G. Chapron, M.-J. Cros, F. Garcia, and R. Sabbadin. MDPtoolbox: A multi-platform toolbox to solve stochastic dynamic programming problems. *Ecography* 37(9):916–920, 2014.

[16] Q. Chen, T. Ayer, and J. Chhatwal. Sensitivity analysis in sequential decision models: A probabilistic approach. *Medical Decision Making* 37(2):243–252, 2017.

[17] J. Chhatwal, O. Alagoz, and E. S. Burnside. Optimal breast biopsy decision-making based on mammographic features and demographic factors. *Operations Research* 58(6):1577–1591, 2010.

[18] M. C. Chou, M. Parlar, and Y. Zhou. Optimal timing to initiate medical treatment for a disease evolving as a semi-Markov process. *Journal of Optimization Theory and Applications* 175(1): 194–217, 2017.

[19] Y. Chow and M. Ghavamzadeh. Algorithms for CVaR optimization in MDPs. Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds. *Advances in Neural Information Processing Systems*, Vol. 27. Curran Associates, Red Hook, NY, 3509–3517, 2014.

[20] P. Clarke, A. Gray, A. Briggs, A. J. Farmer, P. Fenn, R. J. Stevens, D. R. Matthews, I. M. Stratton, and R. R. Holman. 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(10):1747–1759, 2004.

[21] B. A. Craig and P. Sendi. Estimation of the transition matrix of a discrete-time Markov chain. *Health Economics* 11(1):33–42, 2002.

[22] R. B. D'Agostino Jr. Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* 17(19): 2265–2281, 1998.

[23] T. Degris and O. Sigaud. Factored Markov decision processes. O. Sigaud and O. Buffet, eds. *Markov Decision Processes in Artificial Intelligence*. John Wiley & Sons, Hoboken, NJ, 99–126, 2010.

[24] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39(1):1–38, 1977.

[25] J. E. Eckles. Optimum maintenance with incomplete information. *Operations Research* 16(5): 1058–1067, 1968.

[26] C. K. Enders. *Applied Missing Data Analysis*. Guilford Press, New York, 2010.

[27] F. S. Erenay, O. Alagoz, and A. Said. Optimizing colonoscopy screening for colorectal cancer prevention and surveillance. *Manufacturing & Service Operations Management* 16(3):381–400, 2014.

[28] J. A. Filar, D. Krass, and K. W. Ross. Percentile performance criteria for limiting average Markov decision processes. *IEEE Transactions on Automatic Control* 40(1):2–10, 1995.

[29] M. C. Fu. Optimization for simulation: Theory vs. practice. *INFORMS Journal on Computing* 14(3):192–215, 2003.

[30] K. R. Ghani, K. Grigor, D. N. Tulloch, P. R. Bollina, and S. A. McNeill. Trends in reporting Gleason score 1991 to 2001: Changes in the pathologist's practice. *European Urology* 47(2):196–201, 2005.

[31] M. S. Glasgow, B. T. Engel, and B.C. D'Lugoff. A controlled study of a standardized behavioral stepped treatment for hypertension. *Psychosomatic Medicine* 51(1):10–26, 1989.

[32] D. C. Goff, D. M. Lloyd-Jones, G. Bennett, S. Coady, R. B. D'Agostino, R. Gibbons, P. Greenland, D. T. T. Lackland, D. Levy, and C. J. O'Donnell. 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk. *Circulation* 129(25, Suppl. 2):S49–S73 2014.

[33] J. Goh, M. Bayati, S. A. Zenios, S. Singh, and D. Moore. Data uncertainty in Markov chains: Application to cost-effectiveness analyses of medical innovations. *Operations Research.* 66(3): 697–715, 2018.

[34] M. R. Gold, J. E. Siegel, L. B. Russell, and M. C. Weinstein. *Cost-Effectiveness in Health and Medicine.* Oxford University Press, New York, 1996.

[35] R. Gulati, L. Inoue, J. Katcher, W. Hazelton, and R. Etzioni. Calibrating disease progression models using population data: A critical precursor to policy development in cancer control. *Biostatistics* 11(4):707–719, 2010.

[36] M. Hauskrecht and H. Fraser. Planning treatment of ischemic heart disease with partially observable Markov decision processes. *Artificial Intelligence in Medicine* 18(3):221–244, 2000.

[37] M. He, L. Zhao, and W. P. Powel. Approximate dynamic programming algorithms for optimal dosage decisions in controlled ovarian hyperstimulation. *European Journal of Operational Research* 222(2):328–340, 2012.

[38] E. A. M. Heijnsdijk, E. M. Wever, A. Auvinen, J. Hugosson, S. Ciatto, V. Nelen, M. Kwiatkowski, et al. Quality-of-life effects of prostate-specific antigen screening. *New England Journal of Medicine* 367(7):595–605, 2012.

[39] R. A Howard and J. E. Matheson. Risk-sensitive Markov decision processes. *Management Science* 18(7):356–369, 1972.

[40] G. N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research* 30(2):257–280, 2005.

[41] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence* 101(1–2):99–134, 1998.

[42] G. S. Kulkarni, S. M. Alibhai, A. Finelli, N. E. Fleshner, M. A. Jewett, S. R. Lopushinsky, and A. M. Bayoumi. Cost-effectiveness analysis of immediate radical cystectomy versus intravesical Bacillus Calmette-Guerin therapy for high-risk, high-grade (T1G3) bladder cancer. *Cancer* 115(23): 5450–5459, 2009.

[43] K. Kuntz, F. Sainfort, M. Butler, B. Taylor, S. Kulasingam, S. Gregory, E. Mann, J. M. Anderson, and R. L. Kane. Decision and simulation modeling in systematic reviews. Report 11(13)-EHC037-EF, Agency for Healthcare Research and Quality, Rockville, MD, 2013.

[44] M. Kurt, B. T. Denton, A. J. Schaefer, N. D. Shah, and S. A. Smith. The structure of optimal statin initiation policies for patients with type 2 diabetes. *IIE Transaction on Healthcare Engineering.* 1(1):49–65, 2011.

[45] H. Lei, I. Nahum-Shani, K. Lunch, D. Oslin, and S. A. Murphy. A "SMART" design for building individualized treatment sequences. *Annual Review of Clinical Psychology* 8(1):21–48, 2012.

[46] M. L. Litman, A. R. Cassandra, and L. P. Kaelbling. Efficient dynamic-programming updates in partially observable Markov decision processes. Technical Report CS-95-19, Department of Computer Science, Brown University, Providence, RI, 1995.

[47] R. C. Littell, J. Pendergast, and R. Natarajan. Tutorial in biostatistics: Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine* 19(13):1793–1819, 2000.

[48] W. S. Lovejoy. A survey of algorithmic methods for partially observed Markov decision processes. *Annals of Operations Research* 28(1):47–65, 1991.

[49] L. M. Maillart, J. S. Ivy, D. Kathleen, and S. Ransom. Assessing dynamic breast cancer screening policies. *Operations Research* 56(6):1411–1427, 2008.

[50] S. Mannor, D. Simester, P. Sun, and J. N. Tsitsiklis. Bias and variance approximation in value function estimates. *Management Science* 53(2):308–322, 2007.

[51] J. E. Mason and B. T. Denton. A comparison of decision-maker perspectives for optimal cholesterol treatment. *IBM Journal of Research and Development* 56(5):8:1–12, 2012.

[52] G. E. Monohan. A survey of partially observable Markov decision processes: Theory, models, and algorithms. *Management Science* 28(1):1–16, 1982.

[53] S. A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society Series B* 65(2):331–366, 2003.

[54] S. A. Murphy. An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine* 24(10):1455–1481, 2005.

[55] National Institutes of Health. NIH data sharing policy and implementation guidance. https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm, 2003.

[56] A. Nilim and L. E. Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research* 55(5):780–798, 2005.

[57] P. Poupart, K. Kim, and D. Kim. Closing the gap: Improved bounds on optimal POMDP solutions. *Proceedings of the 21st* International *Conference on Automated Planning and Scheduling* (*ICAPS*). AAAI Press, Menlo Park, CA, 2011.

[58] W. B. Powell. *Approximate Dynamic Programming*. John Wiley & Sons, Hoboken, NJ, 2007.

[59] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Hoboken, NJ, 1994.

[60] K. L. Rascati. The $64,000 question—What is a quality-adjusted life year worth? *Clinical Therapeutics* 28(7):1042–1043, 2006.

[61] E. Regnier and S. M. Shechter. State-space size considerations for disease-progression models. *Statistics in Medicine* 32(22):3862–3880, 2013.

[62] J. K. Satia and R. E. Lave Jr. Markovian decision processes with uncertain transition probabilities. *Operations Research* 21(3):728–740, 1973.

[63] A. J. Schaefer, M. D. Bailey, S. M. Shechter, and M. S. Roberts. Modeling medical treatment using Markov decision processes. M. Brandeau, F. Sainfort, and W. Pierskalla, eds. *Handbook of Operations Research/Management Science Applications in Health Care*. Kluwer Academic Publishers, Norwell, MA, 597–616, 2004.

[64] R. F. Serfozo. An equivalence between continuous and discrete time Markov decision processes. *Operations Research* 27(3):616–620, 1979.

[65] S. M. Shechter, M. D. Bailey, A. J. Schaefer, and M. S. Roberts. The optimal time to initiate HIV therapy under ordered health states. *Operations Research* 56(1):20–33, 2008.

[66] U. Siebert, O. Alagoz, A. M. Bayoumi, B. Jahn, D. K. Owens, D. J. Cohen, and K. M. Kuntz. State-transition modeling: A report of the ISPOR-SMDM Modeling Good Research Practices Task Force-3. *Value in Health* 15(6):812–820, 2012.

[67] A. Sinha, M. Rajan, T. Hoerger, and L. Pogach. Costs and consequences associated with newer medications for glycemic control in type 2 diabetes. *Diabetes Care* 33(4):695–700, 2010.

[68] R. D. Smallwood and E. J. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research* 21(5):1071–1088, 1971.

[69] R. L. Smith. Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research* 32(6):1296–1308, 1984.

[70] L. N. Steimle and B. T. Denton. Markov decision processes for screening and treatment of chronic diseases. *Markov Decision Processes in Practice*.Springer International, Cham, Switzerland, 189–222, 2017.

[71] R. J. Stevens, V. Kothari, A. I. Adler, I. M. Stratton, and R. R. Holman. The risk engine: A model for the risk of coronary heart disease in type II diabetes (UKPDS 56). *Clinical Science* 101(6):671–679, 2001.

[72] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*,Vol. 1. MIT Press, Cambridge, MA, 1998.

[73] C. H. Tan and J. C. Hartman. Sensitivity analysis in Markov decision processes with uncertain reward parameters. *Journal of Applied Probability* 48(4):954–967, 2011.

[74] G. W. Torrance. Measurement of health state utilities for economic appraisal: A review. *Journal of Health Economics* 5(1):1–30, 1986.

[75] D. Underwood, J. Zhang, B. T. Denton, N. Shah, and B. Inman. Simulation optimization of PSA threshold based prostate cancer screening policies. *Healthcare Management Science* 15(4):293–309, 2012.

[76] J. Untzer, W. Katon, J. W. Williams, C. M. Callahan, L. Harpole, E. M. Hunkeler, M. Hoffing, et al. Improving primary care for depression in late life: The design of a multi-center randomized trial. *Medical Care* 39(48):785–799, 2001.

[77] C. S. White and H. K. El-deib. Parameter imprecision in finite state, finite action dynamic programs. *Operations Research* 34(1):120–129, 1986.

[78] C. S. White and H. K. El-deib. Markov decision processes with imprecise transition probabilities. *Operations Research* 42(4):739–749, 1994.

[79] W. Wiesemann, D. Kuhn, and B. Rustem. Robust Markov decision processes. *Mathematics of Operations Research* 38(1):153–183, 2013.

[80] Y. Zhang, L. Steimle, and B. T. Denton. Robust Markov decision processes for medical treatment decisions. Working paper, University of Michigan, Ann Arbor, 2017.

[81] J. Zhang, B. T. Denton, H. Balasubramanian, N. Shah, and B. Inman. Optimization of prostate biopsy referral decisions. *Manufacturing & Service Operations Management* 14(4):529–547, 2012.

[82] Y. Zhang, R. G. McCoy, J. E. Mason, S. A. Smith, N. D. Shah, and B. T. Denton. Second-line agents for glycemic control for type 2 diabetes: Are newer agents better? *Diabetes Care* 37(5):1338–1345, 2014.