Optimizing Active Surveillance for Prostate Cancer Using Partially Observable Markov Decision Processes

Weiyu Li[a], Brian T. Denton[a,*], Todd M. Morgan[b]

[a] *Department of Industrial and Operations Engineering, University of Michigan, 1205 Beal Avenue, Ann Arbor, MI 48109, USA*

[b] *Department of Urology, University of Michigan, 1500 E Medical Center Dr SPC 5913, Ann Arbor, MI 48109, USA*

Abstract

We describe a finite-horizon partially observable Markov decision process (POMDP) approach to optimize decisions about whether and when to perform biopsies for patients on active surveillance for prostate cancer. The objective is to minimize a weighted combination of two criteria, the number of biopsies to conduct over a patient's lifetime and the delay in detecting high-risk cancer that warrants more aggressive treatment. Our study also considers the impact of parameter ambiguity caused by variation across models fitted to different clinical studies and variation in the weights attributed to the reward criteria according to patient preferences. We introduce two fast approximation algorithms for the proposed model and describe some properties of optimal policies, including the existence of a control-limit type policy. The numerical results show that our approximations perform well, and we use them to compare the model-based biopsy policies to published guidelines. Although our focus is on prostate cancer active surveillance, there are lessons to be learned for applications to other chronic diseases.

Keywords: Decision process, Medical decision making, OR in medicine, Partially observable Markov decision process, Prostate cancer

1. Introduction

Prostate cancer is the most common cancer in men. The American Cancer Society estimates that almost 250,000 new prostate cancer cases and more than 34,000 deaths will occur in the United States in 2021. Over the last decade, it has become clear that men with low-risk variants of prostate cancer can safely avoid major treatment like surgery and radiation therapy, which may have significant side effects including incontinence and erectile dysfunction (Anandadas et al., 2011). For this reason, active surveillance (AS) has recently become the recommended approach for patients with low-risk prostate cancer. AS involves monitoring patients over time to test for evidence of cancer progression to a high-risk variant of the disease. This allows low-risk cancer patients to enjoy a higher quality of life and possibly avoid treatment altogether (Klotz, 2013).

*Corresponding author

Email addresses: weiyuli@umich.edu (Weiyu Li), btdenton@umich.edu (Brian T. Denton), tomorgan@med.umich.edu (Todd M. Morgan)

AS involves regular testing to monitor a patient's health status. The prostate-specific antigen (PSA) test is a simple blood test in AS that measures PSA amount in the blood serum. High PSA is associated with the presence of prostate cancer. Because the PSA test is very simple with almost no harm, it is commonly used; but high false-positive and negative rates make it unsuitable for AS on its own. Prostate biopsy is the gold standard for AS, which involves sampling tissue with hollow-core needles during an outpatient procedure. Biopsy results are reported using the Gleason grading system, where a Gleason score is assigned by a pathologist to provide a measure of severity of the prostate cancer. Biopsy is much more accurate than the PSA test, but it is still prone to false-negative results if the extracted tissue samples miss the tumor. Biopsies are also very painful, and have potential side effects. Thus, decisions about when to perform biopsies are among the most important decisions for AS.

Unfortunately, there is a lack of consensus among urologists on the best biopsy policy. As one of the first healthcare centers to investigate AS, Johns Hopkins (JH) recommended annual biopsies for patients enrolled in AS (Tosoian et al., 2011). A more recent study conducted by the University of California San Francisco (UCSF) medical center recommends biopsy every two years after diagnosis (Dall'Era et al., 2008). A study at the University of Toronto (U of T) medical center and another European study, the Prostate Cancer Research International Active Surveillance (PRIAS) project, recommend biopsy every three years after diagnosis for patients enrolled in AS (Klotz et al., 2009; Bul et al., 2013).

Deciding the optimal biopsy policy is challenging because: 1) the patient's cancer state is not directly observable due to the inaccuracy of diagnostic tests; 2) cancer progression is a stochastic process; 3) patient preferences about how often to biopsy vary. To address these challenges, we formulated a finite-horizon two-state partially observable Markov decision process (POMDP) model to optimize the biopsy policy for AS using data from the four largest and most well known AS studies referenced above. POMDPs are well suited to this type of optimization problem because the decision-makers (physicians and patients) need to make decisions under conditions of uncertainty about the underlying health state, which progresses stochastically over time, and can only be partially observed from PSA test and biopsy results. Our model seeks to find the optimal biopsy policy that trades off two competing criteria: expected delays in detecting high-risk prostate cancer and the expected number of biopsies.

POMDP models are usually very hard to solve over long time horizons because of the *curse of history* (Pineau et al., 2003). Moreover, the model must be solved multiple times, as we will show in this study, to account for ambiguity in the reward function and the underlying stochastic system associated with each of the cohorts mentioned above. For this reason, we present fast approximation methods that can quickly compute near-optimal solutions. We compare our model-based policies, solved via the approximation methods, to established biopsy guidelines from the literature. We further use inverse optimization to estimate ranges of the implied decision-maker's weights on the two reward criteria. Finally, we combine the results for each of the cohorts to compute a risk-based policy region that partitions the region into three parts: 1) biopsy always recommended; 2) biopsy

never recommended; 3) shared decision-making between the patient and physician is necessary to decide if a biopsy should be performed.

The remainder of this article is organized as follows. In Section 2, we review the relevant literature and describe our contributions to the literature. In Section 3, we formulate our active surveillance POMDP (AS-POMDP) model to optimize the biopsy policy in prostate cancer AS. We describe the exact solution method and two approximation methods for the AS-POMDP model in Section 4, and prove some structural properties of the AS-POMDP model in Section 5. In Section 6, we present the results of optimal policies in our case study. Finally, we conclude in Section 7 and discuss some potential directions for future research.

2. Literature Review

In this section, we briefly review the most relevant literature from the application and methodological perspectives. We then summarize our main contributions in light of the existing literature.

2.1. Applications

Much clinical research has been done in recent years to study prostate cancer AS. Several review articles, including Bastian et al. (2009); Klotz (2010); Dall'Era et al. (2012) and Thomsen et al. (2014), have discussed the clinical implication of prostate cancer AS with the focus on inclusion criteria, biopsy guidelines, patient outcomes, and future research needed. The urology community has largely converged on the appropriateness of AS for patients with low-risk cancer. However, different centers have proposed different AS guidelines, which vary most significantly in the recommended frequency of biopsies (Dall'Era et al., 2008; Klotz et al., 2009; Tosoian et al., 2011; Bul et al., 2013).

Epstein et al. (2012) presented results for predictive risk factors for outcomes of radical prostatectomy, which were instrumental in laying the framework for selection criteria for AS enrollment. More recently Coley et al. (2017) built a Bayesian hierarchical model to estimate the sensitivity and specificity of biopsy, and predict the latent cancer states in the JH study, while assuming no cancer progression. Barnett et al. (2018a) estimated a hidden Markov model to estimate the biopsy accuracy and cancer progression rate implied by observed data in the JH study. They further compared different biopsy guidelines using a simulation model based on the hidden Markov model. A recent study by Li et al. (2020) used a hidden Markov model to estimate the cancer progression rate, biopsy under-sampling error, and PSA distribution in the four largest AS cohorts, including JH hospital, UCSF medical center, U of T medical center, and the PRIAS project. The descriptive models given by this study provide the foundation for the prescriptive POMDP models we present here.

POMDP models have been found to be successful in recent decades for optimizing medical decisions when the health state is not directly observable. POMDP models applied to clinical decision-making include the study of screening based on mammography for breast cancer (Simmons Ivy et al., 2009; Ayer et al., 2012, 2016; Otten et al., 2020), colonoscopy screening for colorectal cancer (Erenay et al., 2014), and liver transplantation decisions in the context of liver disease (Sandıkçı

et al., 2013). The most related work to ours – and the only other work on POMDPs for prostate cancer that we are aware of – is that of Zhang et al. (2012a) and Zhang et al. (2012b), which used a POMDP model to optimize the one-time biopsy policy (i.e., the best timing for biopsy if only one biopsy is allowed) in prostate cancer screening. Their work focused on screening of healthy patients who are asymptomatic, the vast majority of whom receive at most one biopsy. Thus, their model can be viewed as an *optimal stopping time problem*, as opposed to AS that involves a continuous process of sequential follow-up biopsies.

2.2. Methodology Literature

POMDP models were first studied by Åström (1965); Drake (1962) and Smallwood & Sondik (1973), and they have been applied in many contexts including machine maintenance (Ross, 1971), robot navigation (Cassandra et al., 1996), healthcare (Ayer et al., 2012; Zhang et al., 2012a; Erenay et al., 2014), and many others (see Cassandra (1998) for a survey). Smallwood & Sondik (1973) introduced the first exact solution method, referred to as the *one-pass algorithm*, for finite-horizon POMDP models. White (1991) and Littman et al. (1995) later proposed the more efficient *witness algorithm* that achieves computational efficiency through a refined approach for identifying the supporting hyperplanes that define the optimal value function. Zhang & Liu (1996) and Cassandra et al. (1997) introduced the *incremental pruning* algorithm, which has been found to be one of the most efficient exact algorithms for a number of problems. Despite its utility for real world applications, solving POMDP models exactly has been shown to be NP-hard, and in PSPACE (Vlassis et al., 2012), due to the so-called *curse of dimensionality* (Kaelbling et al., 1998) and *curse of history* (Pineau et al., 2003).

Many approximation methods for the POMDP model have been studied in the past several decades. An early survey by Lovejoy (1991) discussed exact solution methods for finite-horizon POMDP models in theory, and their finite-memory and finite-grid approximations. Kaelbling et al. (1998) explored function-approximation methods for approximating the value function of POMDP models. Hauskrecht (2000) surveyed various value-function approximation methods for infinite-horizon problems in the application of agent navigation, analyzed their properties and relations, and also presented some novel approximation methods and refinements of existing methods. Unlike the finite-horizon problem, the infinite-horizon POMDP assumes a stationary (i.e., time-independent) value function, with the discounting factor for future rewards being strictly less than one. Pineau et al. (2003) formally defined the *point-based value iteration* (PBVI) algorithm for infinite-horizon POMDPs, and proved the estimation error is bounded. Spaan & Vlassis (2005) introduced the *Perseus* algorithm, which is closely related to PBVI. A more recent survey of point-based POMDP solvers for infinite-horizon problems was published by Shani et al. (2013). Although there were a number of instances where the existing approximation methods were found to be efficient, the issues of finding the best upper bound of the value function with a guaranteed error bound, especially in finite-horizon problems, can easily become intractable and remains unsolved.

Another topic of interest in the literature has been establishing monotonicity of optimal policies, since such policies can be easier to understand and implement, and maybe easier to solve. Ross

(1971) first investigated the monotonicity of the optimal policy in a two-state production process described by a POMDP model. Albright (1979) proved the sufficient conditions for the monotonicity of the optimal policy in a two-state POMDP with the restriction that the actions are taken to improve, rather than investigate the system. Other works include White (1979), Lovejoy (1987), and Miehling & Teneketzis (2020), which generalized this property to models with more than two states by defining the partial order of the belief.

References	Topic	How it differs from our work
Simmons Ivy et al. (2009)	Screening and treatment for breast cancer	Built a simulation method to evaluate policies using the POMDP model instead of solving for optimal policies.
Ayer et al. (2012)	Screening for breast cancer	The proposed POMDP model was to improve patients' QALYs. The optimal value function was monotone in belief. Only considered exact solution methods, which took more than 55 hours for a single model.
Zhang et al. (2012a)	Screening for prostate cancer	The objective was to improve patients' QALYs. Assumed one-time decision because patients could have at most one biopsy.
Sandıkçı et al. (2013)	Liver transplantation for liver disease	The objective was to improve patients' QALYs. The model had monotone optimal value function. Considered an approximate solution method that incorporated solving an LP at each decision epoch, without a bound on approximation error.
Erenay et al. (2014)	Screening for colorectal cancer	The proposed POMDP model was to improve patients' QALYs
Ayer et al. (2016)	Screening for breast cancer with imperfect adherence behavior	Similar model setting as in Ayer et al. (2012), but incorporates adherence behavior to policies. Only considered exact solution method, which took more than 153 hours for a single model.
Otten et al. (2020)	Post-treatment screening for breast cancer	Similar model setting as in Ayer et al. (2012), but the state space was continuous. Optimized the mammography decision within 10-year follow-up after treatment.

Table 1: Previous work on POMDP models for medical decision-making. All but Zhang et al. (2012a) are in different disease contexts. All have different model structures in terms of states, actions, and optimality equations.

2.3. Contributions to the Literature

Our work makes a novel contribution to the literature in several ways. First, we propose the first model to optimize individualized biopsy policies for prostate cancer AS patients. Our study has a model structure that differs from many previously formulated POMDPs, including −− but not limited to −− those arising in clinical contexts that are summarized in Table 1. We describe the approach we used to formulate this complex clinical problem, which is naturally expressed as a two-state POMDP, and then we evaluate the model using observational data from the four most well-known studies of AS thus far. Second, we analyze the model to provide theoretical insight into the structure of the optimal policy. We also discuss the relationship between model-based dynamic policies that learn based on observed data acquired as patients age, such as our POMDP model provides, and static pre-defined policies that have been recommended in the clinical literature. Third, we provide a means to consider the impact of ambiguity caused by variation across models fitted to different clinical studies as well as variation in the reward criteria. Finally, our work collectively demonstrates the full spectrum of using clinical study data directly to estimate and solve POMDP models for an important medical decision-making problem affecting many men worldwide.

3. POMDP Model Formulation

In this section, we describe the discrete-time finite-horizon POMDP model we use to optimize the policy for prostate cancer AS. As noted in the introduction, the objectives are minimizing 1) expected delay in detection of high-risk prostate cancer; 2) expected number of lifetime biopsies. Clearly, it would be ideal to minimize both of these objectives, but that is not possible because they are competing; therefore, we settle for minimizing a weighted combination of the two criteria. We start by describing two main assumptions that form the basis for POMDP model formulation.

Assumption 1. *Prostate cancer progression can be described using a finite-state (two-state) Markov chain.*

Assumption 1 simplifies the stochastic process of prostate cancer progression to that of a first-order Markov chain. The finite state assumption naturally follows from the binary discrimination of health states on the basis of risk as determined by clinical thresholds using pathology information. We describe additional details about this when we discuss the model formulation.

Assumption 2. *The probability distributions of PSA test and biopsy results are conditionally independent given the current cancer state of the patient.*

Assumption 2 assumes conditional independence for different observations given the state of the process. It is a common assumption in partially observable stochastic models that describes the causal relations between the underlying state and the associated observations, and can be adapted to the study of prostate cancer AS. Assumption 1-2 have been validated in a related study of hidden Markov models for prostate cancer by Li et al. (2020).

With the main modeling assumptions established, we now define the elements of the proposed discrete-time finite-horizon AS-POMDP model. We also describe lesser but still important assumptions as part of the model description.

Decision Epochs. We index $t = 1, ..., T$ as the discrete-time periods (also referred to as decision epochs) at which the decision-maker can choose to biopsy, and the state transitions happen. In the AS-POMDP model, t is an annual epoch and the decision is made at the start of the epoch followed by the state transition. Epochs occur annually because this is an upper bound on the frequency of biopsies according to clinical guidelines, i.e., no guideline suggests biopsies more frequently than annually. Epoch $t = 1$ denotes the time of diagnosis and enrollment in AS, and epoch $t = T$ is the recommended stopping time for AS among patients who survive until age T, which is typically age 75 according to clinical guidelines due to increases in competing causes of death.

States. The set of states, S, contains two states: 1) low-risk prostate cancer state (LR); 2) high-risk prostate cancer state (HR). In reality, there are numerous health states defined by risk factors, including PSA and pathology from biopsies; however, urologists differentiate these states into two groups (LR and HR) to align clinical risk with treatment choices. Patients who are known to be in the LR state should continue AS, while those in the HR state should be treated (e.g., surgery or radiation therapy). We use s_t to denote the state of the system at time t for $t = 1, ..., T$.

Actions. The set of available actions, A, contains two elements: 1) defer biopsy; 2) conduct biopsy. As the PSA test is always done by default according to standard clinical practice, the critical decision at each decision epoch is whether or not to conduct a biopsy. Note that in prostate AS, the action of conducting biopsy is to investigate, rather than to improve, the patient's cancer state. In other words, conducting a biopsy does not affect the stochastic process of cancer progression (unless the patient leaves AS for treatment because of a biopsy Gleason score upgrading defined later).

Transition Probabilities. At each decision epoch, the system undergoes state transitions according to transition probability P defined as follows:

$$P(i,j) := \mathbb{P}(s_{t+1} = j | s_t = i),\ \forall i,j \in S,\ \forall t = 1, ..., T-1.$$

In our AS-POMDP model, the state can only progress from LR cancer to HR cancer, so that we use p to denote this annual progression rate.

Observations. At each decision epoch, after an action is taken, the PSA test and biopsy result (if conducted) will be observed. We denote O as the set of all possible observations, and $o_t \in O$ as the observation at time t for $t = 1, ..., T$. By Assumption 2, at any decision epoch, given the state of the system, the observations of PSA test and biopsy are mutually independent. So, $O = O_{\text{PSA}} \times O_{\text{Biopsy}}$, where O_{PSA} is the observation space of the PSA test, and O_{Biopsy} is the observation space of the biopsy. We discretize the space of the measurement of PSA levels into three intervals, according to the widely used PSA cutoffs in clinical studies (Hoffman, 2011), so that O_{PSA} has three elements: $I_1 = [0, 4]$, $I_2 = (4, 10]$, and $I_3 = (10, \infty)$ (ng/mL). For biopsy, the elements in O_{Biopsy} are Gleason score upgrading (biopsy Gleason score greater or equal to 7), Gleason score not upgrading (biopsy Gleason score less or equal to 6), and null observation (biopsy not conducted). Such definition is based on the fact that the inclusion criteria for AS in all four study centers considered in this paper require the biopsy Gleason score to be less or equal to 6 (Li et al., 2020). We now state the third assumption of the AS-POMDP model formulation as follows.

Assumption 3. *Patients leave AS immediately when a biopsy Gleason score upgrading is observed.*

Assumption 3 is reasonable because Gleason score upgrading is a common criterion for dropping from prostate cancer AS in practice, as well as in the studies used to parameterize and test our model. In some cases, the decision is nuanced, requiring a shared decision-making approach between the patient and physician because of considerations of age, comorbidities, and the patient's personal preferences. However, our AS-POMDP model assumes such patients leave the system and receive care that is specialized to their personal situation with the guidance of a urologist.

Observation Probabilities. The observation probability is defined as the probability of observing certain output given the state of the system and the action taken. In the AS-POMDP model, the

observation probabilities $\mathbb{P}(o|s,a)$ for all $a \in A$ and $o = (x,y) \in O = O_{\text{PSA}} \times O_{\text{Biopsy}}$ are given by

$$\mathbb{P}((x,y)|s,a) = \begin{cases} q^{\text{LR}}(I_i), & a = \text{Defer Biopsy},\ s = \text{LR},\ y = \text{Null},\ x \in I_i,\ \forall i; \\ q^{\text{HR}}(I_i), & a = \text{Defer Biopsy},\ s = \text{HR},\ y = \text{Null},\ x \in I_i,\ \forall i; \\ q^{\text{LR}}(I_i), & a = \text{Conduct Biopsy},\ s = \text{LR},\ y = \text{Not Upgrading},\ x \in I_i,\ \forall i; \\ \gamma q^{\text{HR}}(I_i), & a = \text{Conduct Biopsy},\ s = \text{HR},\ y = \text{Not Upgrading},\ x \in I_i,\ \forall i; \\ (1-\gamma) q^{\text{HR}}(I_i), & a = \text{Conduct Biopsy},\ s = \text{HR},\ y = \text{Upgrading},\ x \in I_i,\ \forall i; \\ 0, & \text{otherwise}, \end{cases}$$

where q^{LR} and q^{HR} are probability mass functions of PSA in the LR and HR cancer states, and γ is the false-negative rate (1 - sensitivity) of biopsy defined as the probability of observing Gleason score not upgrading while in HR cancer state.

Here we assume that biopsies have perfect specificity, i.e., the probability of observing a Gleason score upgrading when in LR cancer state is zero. This is because biopsies involve sampling of prostate tissue, and thus sometimes may miss the tumor; however when the tumor is sampled, the probability that it is identified by a qualified pathologist in nearly 1.

Reward Function. We let $r(s,a,o)$ denote the reward function when the system is in state $s \in \mathcal{S}$, action $a \in A$ is taken, and output $o = (x,y) \in O$ is observed at each decision epoch, which is given by

$$r(s,a,(x,y)) = \begin{cases} 0, & a = \text{Defer Biopsy},\ s = \text{LR}; \\ \theta, & a = \text{Defer Biopsy},\ s = \text{HR}; \\ \eta, & a = \text{Conduct Biopsy},\ s = \text{LR},\ y = \text{Not Upgrading}; \\ \eta, & a = \text{Conduct Biopsy},\ s = \text{HR},\ y = \text{Upgrading}; \\ \theta + \eta, & a = \text{Conduct Biopsy},\ s = \text{HR},\ y = \text{Not Upgrading}; \\ \text{Not Defined}, & \text{otherwise}, \end{cases}$$

where θ and η are non-positive scalars that denote the negative reward (cost) of one-year delayed detection of high-risk cancer and the burden of a biopsy, respectively. In the AS-POMDP model, we seek to minimize a weighted sum of the expected number of biopsies and years in late detection to cancer progression, so these are negative "rewards." Note that θ and η are pre-determined scalars that reveal the decision-maker's consideration of the two events. We set $\theta + \eta = -1$, so that varying θ and η allows computing the optimal policy for different patient preferences for the two criteria.

Figure 1 illustrates the stochastic control process of the proposed AS-POMDP model. At the beginning of each decision epoch, the decision-maker can choose the test action of whether to defer and conduct biopsy. Then, the test outcome is observed, which provides partial information about the underlying cancer state. Given the chosen action and test outcome, an immediate negative reward is assigned to the patient, which comes from the burden of test action and/or the penalty of failing to detect a cancer progression to the HR state, if there was one. If the biopsy result shows

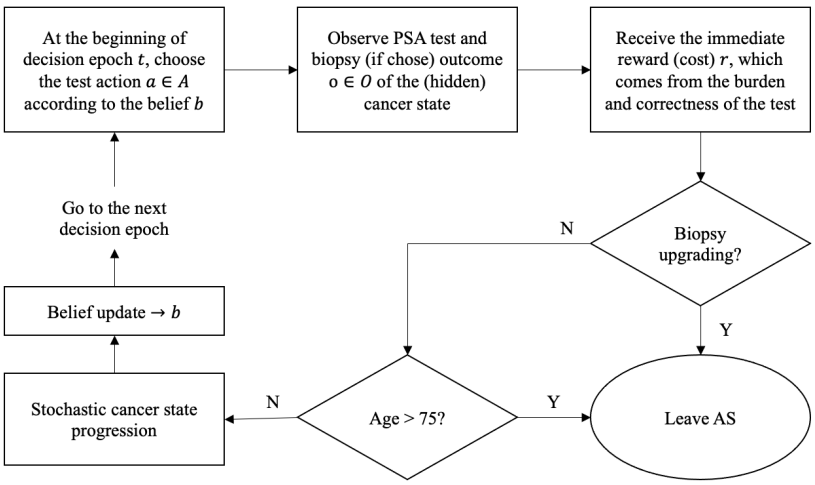

Figure 1: The stochastic control process of prostate cancer AS described by the proposed AS-POMDP model

Gleason score upgrading, then the patient will leave AS immediately. For the case that shows no upgrading or null biopsy, if the patient is older than age 75, he will also leave AS.

Belief state. We use b_t to denote the belief, i.e., the probability distribution, over the set of states, S, at the beginning of decision epoch t. In the AS-POMDP model, since there are only two states in S, the belief b_t only has one degree-of-freedom and can be represented by the probability of being in the HR cancer state, with $1-b_t$ being the probability of being in the LR cancer state. In particular, for the starting time $t = 1$, b_1 is the probability that the patient who enters AS (because of being diagnosed with LR cancer) is actually in HR cancer state, i.e., misclassification error at diagnosis. The belief b_t at epoch t is well-known to be a sufficient statistic for the past sequence of actions and observations before time t. We use Λ to denote the Bayes updating formula from time t to $t+1$, i.e.,

$$b_{t+1} = \Lambda(b_t|a, o),\ 1 \le t \le T-1, \tag{1}$$

if action a is taken and output o is observed. The exact expression of Λ is given in the next section. Notice that sometimes we may drop the subscript of b_t when it is treated as the argument of the value function defined later.

Policy. A policy $\pi = (\pi_1, ..., \pi_T)$ is defined as a set of functions from the belief space to the action space, where π_t specifies the actions to take for all possible belief states at decision epoch $t = 1, ..., T$.

Value Function. Given a policy π, we define the expected cumulative reward starting from time t_0 until the end of the time horizon T as:

$$V_{t_0}^{\pi}(b) := \mathbb{E}^{\pi}[\sum_{t=t_0}^{T} r(s_t, a_t, o_t)|b],\ \forall b, \forall t_0,$$

where the expectation is taken over all possible state, action, and observation trajectories following the policy π. For a fixed π and t_0, $V_{t_0}^{\pi}(b)$ is a function of the belief state b. Note that in the

AS-POMDP model, if the patient leaves AS because of a Gleason score upgrading before time T, then the process will stop with no future reward.

As discussed in Smallwood & Sondik (1973), the POMDP model can be viewed as a continuous-state Markov decision process model with the state space being the space of all possible belief states. It follows immediately that there exists an optimal policy π^* that is deterministic and Markovian with respect to the belief, which maximizes the expected cumulative rewards at any time t:

$$V_t(b) := V_t^{\pi^*}(b) = \max_{\pi} V_t^{\pi}(b), \ \forall b,$$

which can be computed using the following optimality equations:

$$V_t(b) = \max_{a \in A} \{ \sum_{s \in S} b(s) r(s,a) + \sum_{o \in O} \mathbb{P}(o|b,a) V_{t+1}(\Lambda(b|a,o)) \}, \ \forall b, \ \forall t,$$

with the boundary condition

$$V_T(b) = \max_{a \in A} \sum_{s \in S} b(s) r(s,a), \ \forall b,$$

where $r(s,a) = \sum_{o \in O} \mathbb{P}(o|s,a) r(s,a,o)$ is the expected immediate reward when the system is in state s and action a is taken, and $\mathbb{P}(o|b,a) = \sum_{s \in S} b(s) \mathbb{P}(o|s,a)$ is the probability of observing o when the belief is b and action a is taken. By Assumption 3, in our AS-POMDP model, since the patient will leave AS upon receiving a biopsy that shows Gleason score upgrading, the optimality equation for $t < T$ should be modified as

$$V_t(b) = \max_{a \in A} \{ \sum_{s \in S} b(s) r(s,a) + \sum_{o \in O'} \mathbb{P}(o|b,a) V_{t+1}(\Lambda(b|a,o)) \}, \ \forall b, \ \forall t, \tag{2}$$

where $O' = O_{\text{PSA}} \times \{\text{Not Upgrading}, \text{Null}\}$ is a subset of O. Solving the optimality equations yields the optimal policy $\pi^* = (\pi_1^*, ..., \pi_T^*)$ as follows

$$\pi_t^*(b) := \arg\max_{a \in A} \{ \sum_{s \in S} b(s) r(s,a) + \sum_{o \in O'} \mathbb{P}(o|b_t,a) V_{t+1}(\Lambda(b|a,o)) \}, \ \forall b, \ \forall t < T,$$

and

$$\pi_T^*(b) := \arg\max_{a \in A} \sum_{s \in S} b(s) r(s,a), \ \forall b.$$

4. Solution Methods

In this section, we describe the approach we used to solve the AS-POMDP model formulated in Section 3. We start by describing an exact solution method, the classical *one-pass algorithm* of Smallwood & Sondik (1973), to set the foundation for describing our approach. Unfortunately, the one-pass algorithm is impractical for the AS-POMDP model, as the number of non-dominated α-vectors is growing exponentially in the size of the observation space at each time period. Because of the long time horizon and the fact that we intend to solve a number of different AS-POMDP model instances with different choices of model parameters, fast approximation methods are preferred over

the exact method (which took more than 24 hours for a single set of model parameters using an Intel Core i7 2.6 GHz processor with 16 GB RAM). Therefore, we study two approximation methods that give lower and upper bounds on the optimal value function, with bounded worst-case approximation errors. We further show in the numerical results that the gaps between the lower and upper bounds are very small so that our approximate solutions are accurate enough to be trusted.

4.1. Exact Solution Method

As shown in Smallwood & Sondik (1973), the optimal value function $V_t(b)$ is piece-wise linear and convex in b, and can be written as

$$V_t(b) = \max_{\alpha \in \mathcal{A}_t} \alpha \cdot b, \ \forall b, \ \forall t,$$

where $\mathcal{A}_t$ is a set of linear functions, referred to as α-*vectors*, which be calculated by backward induction. Further, each α-vector in $\mathcal{A}_t$ corresponds to a decision tree that specifies the choices of action for all possible observations at each of the future decision epochs (see Kaelbling et al. (1998) for more details). It is easy to see that this property is also true in the AS-POMDP model, although equation (2) omits a part of *value-to-go* (which is linear in the belief) if the patient leaves the system before the end of AS due to observing a Gleason score upgrading. With this property, the AS-POMDP model can be solved by finding the set of α-vectors, $\mathcal{A}_t$, at each decision epoch t.

In our AS-POMDP model, since there are only two states, the belief can be represented by a scalar. We let b denote the belief in the high-risk cancer state, and thus the belief in the low-risk cancer state is $1 - b$. Further, in the AS-POMDP model, each α-vector is a line, and can be determined by any two points on the line. For convenience, we use a vector, $(l(0), l(1))$, to represent the linear function $l(b)$, where $l(0)$ and $l(1)$ are the values of l at points $b = 0$ and $b = 1$, respectively. For models with more than two states, it is easy to generalize our results by using the extreme points of the belief simplex to represent α-vectors.

Starting with the boundary condition, the optimal value function at time T can be written as

$$V_T(b) = \max_{a \in A}\{(1-b)r(s_1, a) + br(s_2, a)\}, \forall b,$$

where $r(s_i, a) = \sum_y r(s_i, a, y)$ for $i = 1, 2$. So,

$$\mathcal{A}_T = \{(r(s1, a_1), r(s_2, a_1)), (r(s_1, a_2), r(s_2, a_2)\}.$$

Now, given the set of α-vectors, $\mathcal{A}_{t+1}$ at time $t+1$, to derive the set of α-vectors, $\mathcal{A}_t$ at time t by way of backward induction, we can, for each α-vector in $\mathcal{A}_t$, find its values of at $b = 0$ and $b = 1$. In our AS-POMDP model, for each decision epoch, the belief update $b_{t+1} = \Lambda(b_t|a, o)$ is realized in two steps as follows,

$$b_t \xrightarrow{obs.} \tilde{b}_t \xrightarrow{trans.} b_{t+1}.$$

Specifically, suppose at time t, action a was taken and we observed o, then

$$\tilde{b}_t = \frac{\mathbb{P}(o|s_2, a)}{\mathbb{P}(o|b_t, a)} b_t,$$

and

$$b_{t+1} = \Lambda(b_t|a,o) = \tilde{b}_t + p(1-\tilde{b}_t) = \frac{1}{\mathbb{P}(o|b_t,a)}((1-p)\mathbb{P}(o|s_2,a)b_t + p\mathbb{P}(o|b_t,a)).$$

Then, by the optimality equations (2), for a specific action a, each α-vector $\alpha_t = (\alpha_t(0), \alpha_t(1))$ at time t can be be represented by

$$\alpha_t(0) = r(s_1,a) + \sum_{o\in O'} \mathbb{P}(o|0,a)\alpha_{t+1,o}(\Lambda(0|a,o)) = r(s_1,a) + \sum_{o\in O'} \mathbb{P}(o|0,a)\alpha_{t+1,o}(p)$$

and

$$\alpha_t(1) = r(s_2,a) + \sum_{o\in O'} \mathbb{P}(o|1,a)\alpha_{t+1,o}(\Lambda(1|a,o)) = r(s_2,a) + \sum_{o\in O'} \mathbb{P}(o|1,a)\alpha_{t+1,o}(1)$$

for a specific set of choices of $\alpha_{t+1,o} \in \mathcal{A}_{t+1}$ for all $o \in O'$. Enumerating all such sets of choices of α-vectors at time $t+1$ and actions gives all α-vectors at time t, which we denote as $\tilde{A}_t$; however, some of the α-vectors in $\tilde{A}_t$ can be dominated by the others and thus can be *pruned* by solving a linear program (Smallwood & Sondik (1973)). Littman et al. (1995) and Zhang & Liu (1996) proposed the witness and incremental pruning algorithms that improve the pruning procedure and generate the minimal set of non-dominated α-vectors $\mathcal{A}_t$ at time t.

We let H denote the operator for the backward induction and pruning steps of V_t from V_{t+1} in the one-pass algorithm described above, and write the optimality equations as

$$V_t = HV_{t+1}, \quad t = T-1, ..., 1.$$

4.2. Point-based Approximation Method

The point-based approximation method is well-suited here, because instead of finding the set of all dominated α-vectors at each decision epoch, it only evaluates the value function at a set of sampled belief points to get an estimate of the value function. And by controlling the number of the sampled belief points, it limits the number of α-vectors to keep at each decision epoch. Different types of point-based value function approximation methods have been carefully studied in the surveys of Hauskrecht (2000), Pineau et al. (2003), and Shani et al. (2013) for infinite-horizon POMDPs, where the value function was assumed to be stationary (i.e., independent with time). We generalize their approach to our finite horizon non-stationary AS-POMDP model. In this section, we let B_t denote the sampled belief points at decision epoch t, and provide the methods for finding the lower and upper bounds of the value functions based on B_t for all $t = 1, ..., T$.

4.2.1. Lower Bound

At each decision epoch t, since the optimal value function can be written as the maximum of a set of linear functions in $\mathcal{A}_t$, a natural way to find a lower bound of V_t is to use a subset of $\mathcal{A}_t$. Starting from the optimal value function at the next decision epoch V_{t+1} and associated α-vectors, $\mathcal{A}_{t+1}$, we first derive the set of all (dominated and non-dominated) α-vectors $\tilde{\mathcal{A}}_t$ following the steps described in Section 4.1. Then, at each belief point, $b \in B_t$, we identify the supporting α-vectors in $\tilde{\mathcal{A}}_t$, resulting in $|B_t|$ α-vectors being selected from $\tilde{\mathcal{A}}_t$. We denote the set of selected α-vectors

as $\hat{\mathcal{A}}_t$. Thus, at each decision epoch t, $\hat{V}_t$ defined as follows gives a lower bound of the true value function V_t:

$$\hat{V}_t(b) := \max_{\alpha \in \hat{\mathcal{A}}_t} \alpha \cdot b, \ \forall b.$$

The details of the lower bound approximation method is described in Algorithm 1.

Algorithm 1: Algorithm for approximate backward induction with operator L^B.

Input : V_{t+1}, B
Output: $\hat{V}_t$
Initialize $\hat{\mathcal{A}}_t$ as a empty set;
Let $\mathcal{A}_{t+1}$ as the set of α-vectors defining V_{t+1};
Find the set of all α-vectors at time t, $\tilde{\mathcal{A}}_t$ using $\mathcal{A}_{t+1}$ and backward induction;
for $b \in B$ **do**
  $\alpha_b \leftarrow \arg\max_{\alpha \in \tilde{\mathcal{A}}_t} \alpha \cdot b$;
  add α_b in $\hat{\mathcal{A}}_t$;
end
Define $\hat{V}_t(b) := \max_{\alpha \in \hat{\mathcal{A}}_t} \alpha \cdot b, \ \forall b$.

We let operator L^B denote approximate backward induction steps described in Algorithm 1. Note that L^B needs not to start from the exact optimal value function at the next decision epoch. If we start from any subset of $\mathcal{A}_{t+1}$, and the corresponding lower bound on V_{t+1}, then L^B will also provide a lower bound on V_t because the resulting $\hat{A}_t$ is always a subset of $\mathcal{A}_t$. In particular, if we start from the boundary condition V_T, with the sample belief sets B_t for all $t = 1, ..., T-1$, then

$$\hat{V}_t = L^{B_t} L^{B_{t+1}} ... L^{B_{T-1}}(V_T), \forall t = 1, .., T-1 \tag{3}$$

is always a lower bound of V_t. The following theorem gives the error bound between $\hat{V}_t$ and V_t for each t, whose proof utilizes the triangle inequality and Holder's inequality and is adapted from Theorem 3.1 of Pineau et al. (2003). The proof of the Theorem 1 is in the Appendix.

Theorem 1. *Given the grids of the belief space at each decision epoch $B_t \subset [0,1]^{|S|}$ for all t, the error between the optimal value function V_t and approximated value function $\hat{V}_t$ given by (3) satisfies*

$$||V_t - \hat{V}_t||_\infty \leq \frac{(T-t)(T-t+1)}{2} ||r_{\max} - r_{\min}||_\infty \delta,$$

where

$$r_{\max}(s) := \max_{a \in A} \sum_{o \in O} \mathbb{P}(o|s,a) r(s,a,o), \ r_{\min}(s) := \min_{a \in A} \sum_{o \in O} \mathbb{P}(o|s,a) r(s,a,o), \ \forall s \in S,$$

and

$$\delta := \max_t \max_{b \in [0,1]^{|\mathcal{S}|}} \min_{b' \in B_t} ||b' - b||_1.$$

The bound in Theorem 1 tends to zero as $\delta \to 0$.

4.2.2. Upper Bound

Approaches to upper bound the optimal value function often involve solving many linear programs (Hauskrecht, 2000). Fortunately, for a two-state POMDP model such as the AS-POMDP model, the solution of the linear program can be given directly, which can further accelerate approximate backwards induction for our two-stage AS-POMDP model. At each decision epoch t, given the set of α-vectors A_{t+1} that defines V_{t+1} in the next decision epoch, to find the upper bound of V_t, we use the linear interpolation of the sampled belief points and their values. Specifically, given the sampled belief set B_t, for each $b \in B_t$, we first calculate $u_t(b) := V_t(b)$ using the optimality equation. Then, as long as B contains the extreme points $b = 0$ and $b = 1$, for any belief point $b' \in [0, 1]$, the solution of the following linear program will give the best linear interpolation for $V_t(b')$:

$$\begin{aligned} \bar{V}_t(b') := \min_{\lambda} \quad & \textstyle\sum_{b\in B} \lambda_b u_t(b) & & \\ \text{s.t.} \quad & \textstyle\sum_{b\in B} \lambda_b &= 1, & \\ & \lambda_b &\geq 0, & \forall b \in B \\ & \textstyle\sum_{b\in B} \lambda_b b &= b'. & \end{aligned}$$

Further, $\bar{V}_t$ is an upper bound of V_t.

For a two-state POMDP model such as ours, the following results show that the optimal solution to the linear program is trivial, so that an upper bound of V_t can be obtained without resorting to solving linear programs.

Proposition 1. *In a two-state POMDP model, at time t, write the set of the sample belief point B_t as*

$$B_t = \{b^1, b^2, ..., b^{|B|}\}$$

such that $0 = b^1 < b^2 < ... < b^{|B|} = 1$. Then, for every $b' \in [0, 1]$ such that $b^i \leq b' < b^{i+1}$, the optimal solution of the above linear program has only two variables λ_{b^i} and $\lambda_{b^{i+1}}$ being non-zero, and all others being zero.

Proof. Notice that the linear program has $|B|$ decision variables λ_b for $b \in B$ and $|B|+2$ constraints. Then, the extreme point of the polyhedron defined by the constrains should satisfy $|B|-2$ equations of $\lambda_b = 0$ for $b \in B$.

Now, for $b' \in [0, 1]$ such that $b^i \leq b' < b^{i+1}$, suppose the extreme value $\bar{V}_t(b')$ is achieved with two λ_{b^j} and λ_{b^k} being non-zero, and all other decision variables being zero. Notice that to satisfy the first and last constraints, we can assume $b^j \leq b^i$ and $b^k \geq b^{i+1}$ without the loss of generality. Then, since V is convex, at b_t, the convex combination of b^j and b^k is greater than b^j and b^{i+1}, and the convex combination of b^j and b^{i+1} is greater than b^i and b^{i+1}. So, the optimal value $\bar{V}_t(b')$ is achieved with only λ_{b^i} and $\lambda_{b^{i+1}}$ being non-zero, and all other decision variables being zero. □

The above proposition shows that for belief points between b^i and b^{i+1}, $\bar{V}_t$ is defined by the line determined by two points $(b^i, u_t(b^i))$ and $(b^{i+1}, u_t(b^{i+1}))$, for all $i = 1, ..., |B| - 1$.The next proposition gives an expression of $\bar{V}_t$.

Proposition 2. *In a two-state POMDP model, at decision epoch t, denote β_i as the linear function determined by $(b^i, u_t(b^i))$ and $(b^{i+1}, u_t(b^{i+1}))$, for all $i = 1, ..., |B| - 1$, and let $\mathcal{B}$ be the set of all such linear functions:*

$$\mathcal{B} := \{\beta_1, ..., \beta_{|B|-1}\}.$$

Then,

$$\bar{V}_t(b_t) = \max_{\beta \in \mathcal{B}} \beta \cdot b_t, \ \forall b_t.$$

Proof. For each $i = 1, ..., |B| - 1$, since β^i is a line determined by $(b^i, u_t(b^i))$ and $(b^{i+1}, u_t(b^{i+1}))$, and V_t is convex, then for $b \in (b^i, b^{i+1})$, $V_t(b) \leq \beta^i \cdot b$; for $b = b^i$ or $b = b^{i+1}$, $V_t(b) = \beta_i \cdot b$; and for $b \notin [b^i, b^{i+1}]$, $V_t(b) \geq \beta_i \cdot b$. By Proposition 1, for $b \in [b^i, b^{i+1}]$, $\bar{V}_t(b) = \beta_i \cdot b = max_{\beta \in \mathcal{B}} \beta \cdot b_t$. □

Algorithm 2 describes the steps for deriving the upper bound of the value function at each decision epoch by approximate backward induction. For convenience, we use operator U^B to denote Algorithm 2 for a given B. Note that the input of U^B can also be any upper bound of V_{t+1}, and the output $\bar{V}_t$ is always an upper bound of V_t because $u_t(b)$ is always greater than $V_t(b)$ for all $b \in B$. In particular, if we start from the boundary condition V_T, with the sample belief sets B_t for all $t = 1, ..., T - 1$, then

$$\bar{V}_t = U^{B_t} U^{B_{t+1}} ... U^{B_{T-1}}(V_T), \forall t = 1, .., T - 1. \tag{4}$$

is always an upper bound of V_t. The next theorem gives the error bound between $\bar{V}_t$ and V_t for each t. The proof of the Theorem is in the Appendix.

Algorithm 2: Algorithm for approximated backward induction U^B.

Input : V_{t+1}, B
Output: $\bar{V}_t$
Initialize $\mathcal{B}$ as a empty set;
Write $\mathcal{B} = \{b_1, ..., b_{|B|}\}$ such that $0 = b_1 < ... < b_{|B|=1}$;
for $b \in B$ **do**
  Calculate $u_t(b) := \max_a \{b \cdot r^a + \sum_o \mathbb{P}(o|b, a) V_{t+1}(U(b|a, o))\}$;
end
for $i = 1$ **to** $|B| - 1$ **do**
  Let β_i be the line determined by two points $(b_i, u_t(b_i))$ and $(b_{i+1}, u_t(b_{i+1}))$;
  Add β_i in $\mathcal{B}$;
end
Define $\bar{V}_t(b) = \max_{\beta \in \mathcal{B}} \beta \cdot b$ for all $b \in [0, 1]$;

Theorem 2. *Given the grids of the belief space $B_t \subset [0, 1]^{|S|}$ at each decision epoch t, the error between the optimal value function V_t and approximated value function $\bar{V}_t$ given by (4) satisfies*

$$||V_t - \bar{V}_t||_\infty \leq \frac{(T - t)(T - t + 1)}{2} ||r_{\max} - r_{\min}||_\infty \delta, \ \forall t \leq T$$

where $r_{\max}$, $r_{\min}$, and δ are defined the same as in Theorem 1.

Remark 1. *Later in Section 6, we show that the actual observed differences between the lower and upper bounds of the value functions in AS-POMDP all models were much smaller than the error bound given by Theorem 1 and 2. This is because in the AS-POMDP model, a patient will leave AS for treatment immediately after observing a Gleason score upgrading, with no future cost. As a result, the expected value-to-go for conducting biopsy, as shown in the optimality equation (2), is shrunk by γ (biopsy false-negative rate). This further makes the error of the approximate value function much smaller than the worst case described in the proof of Theorem 1 and 2. However, since we do not know in advance what is the optimal action at each decision epoch, it is very difficult to improve the error bound. In the extreme case (e.g., always defer biopsy), it is possible that the error bound in Theorem 1 or 2 is achieved with equality. On the other hand, the results in Section 6 show that the proposed approximation methods work very well for the AS-POMDP model.*

5. Structural Properties

In this section, we discuss some structural properties of the proposed AS-POMDP model to provide some insight into the results we present in Section 6.

5.1. Control-limit Type Policy

In Section 6 we will see the solution to the AS-POMDP model is a control-limit type policy, i.e., there is a threshold on the element of the belief vector that represents the probability of being in the high-risk state, below which it is optimal to defer biopsy, and above which it is optimal to conduct biopsy. There are many prior works that have discussed the existence of a control-limit type policy in a POMDP model. For example, White (1979) proved that the optimal replacement policy for the machine maintenance problem is a control-limit type policy. However, one of the distinctions of our model compared to the prior works is that our goal is to inspect and classify the system state (low-risk or high-risk cancer) rather than sequential system improvement, so that the optimal value function in our model is not monotone w.r.t. the belief anymore.

As in Section 4, we denote the set of non-dominated α-vectors at decision epoch t as $\mathcal{A}_t = \{\alpha_1, ..., \alpha_n\}$, and write the optimal value function at time t as

$$V_t(b) = \max_{\alpha_i \in \mathcal{A}} \alpha_i(b), \ \forall b.$$

Then, it is easy to see that V_t has $n-1$ inflection points on $(0, 1)$. The following lemma establishes a useful relationship among the positions of these $n-1$ inflection points, and the relationship between the slopes and endpoints of the non-dominated α-vectors.

Lemma 1. *For $\mathcal{A}_t = \{\alpha_1, ..., \alpha_n\}$, assume that slope$(\alpha_1)$ < slope(α_2) < ... < slope(α_n). Let the positions of the inflection points of V_t to be $b_1 < b_2 < ... < b_{n-1}$. Then, $(b_i, V_t(b_i))$ must be the intersection of α_i and α_{i+1}, $i = 1, ..., n-1$. Further,*

$$\alpha_1(0) > \alpha_2(0) > ... > \alpha_n(0),$$

and

$$\alpha_1(1) < \alpha_2(1) < ... < \alpha_n(1).$$

Proof. We prove the first part by contradiction. Suppose $(b_j, V_t(b_j))$ is the first inflection point of v_t such that it is not the intersection of α_j and α_{j+1}. Then, it should the intersection of α_j and α_k with $k > j+1$. So, $V_t(b) = \alpha_k(b)$ on $b \in (b_j, b_{j+1})$. Since α_{j+1} is not dominated, there must exists some $b_l \geq b_{j+1}$, such that $V_t(b) = \alpha_{j+1}(b)$ on $b \in (b_l, b_{l+1})$. Then, the slope of $V_t(b)$ is not increasing, which contradicts the convexity of V_t.

Now, choose $\beta = (\beta(1), \beta(2)) \in \mathcal{A}_t$ such that $\beta(2) = \min_{\alpha_i \in \mathcal{A}_t} \alpha_i(2)$. Then, it must be true that $\beta(1) = \max_{\alpha_i \in \mathcal{A}_t} \alpha_i(1)$; otherwise, β must be dominated by some α-vectors in $\mathcal{A}_t$. It is easy to see that the slope of β is the smallest in $\mathcal{A}_t$. So, $\beta = \alpha_1$. Remove α_1 from $\mathcal{A}_t$ and repeat the same steps until there is no element in $\mathcal{A}_t$ completes the proof. □

We now leverage the above lemma to provide a sufficient and necessary condition for the existence of a control-limit type policy in a two-dimension POMDP model.

Lemma 2. *For any time t, denote the set of non-dominated α-vectors at time t as $\mathcal{A}_t = \{\alpha_1, ..., \alpha_n\}$. Further, let $\mathcal{A}_t^1 = \{\alpha_1, ..., \alpha_m\}$ be the α-vectors corresponding to action "defer biopsy", and $\mathcal{A}_t^2 = \{\alpha_{m+1}, ..., \alpha_n\}$ be the α-vectors corresponding to action "conduct biopsy". We say $\mathcal{A}_t^1$ and $\mathcal{A}_t^2$ are separable at some $b \in [0,1]$, if at b all values of the α-vectors in $\mathcal{A}_t^1$ are greater or smaller than all values of the α-vectors in $\mathcal{A}_t^2$. Then, the optimal policy at time t is a control-limit type policy if and only if $\mathcal{A}_t^1$ and $\mathcal{A}_t^2$ are separable at $b = 0$, or equivalently, $\mathcal{A}_t^1$ and $\mathcal{A}_t^2$ are separable at $b = 1$.*

Proof. The existence of a control-limit type policy is equivalent to the existence of an inflection point $\bar{b}$ of $v_t(b)$, such that for $b \leq \bar{b}$, $V_t(b)$ is composed of the α-vectors in $\mathcal{A}_t^1$ and for $b > \bar{b}$, $V_t(b)$ is composed of the α-vectors in $\mathcal{A}_t^2$; Further, if there exists an inflection point $\bar{b}$ of $V_t(b)$, such that for $b < \bar{b}$, then the inflection points of $V_t(b)$ are the intersections between the α-vectors in $\mathcal{A}_t^1$; and for $b > \bar{b}$, the inflection points of $v_t(b)$ are the intersections between the α-vectors in $\mathcal{A}_t^2$. According to Lemma 1, the inflection points following the sequence of the slopes of the α-vectors, and the order of the slopes of the α-vectors is equivalent to the order of the values of the α-vectors at either endpoint. □

Focusing on our AS-POMDP model specifically, we let γ denote the false-negative rate of the biopsy, and note that the expected immediate reward for action "defer biopsy" at $b = 1$ is θ and the expected immediate reward for action "conduct biopsy" at $b = 1$ is $\eta + \gamma\theta$ $(= -1 - \theta + \gamma\theta)$. We only consider the case where $\eta + \gamma\theta$ is greater than θ, i.e., "conduct biopsy" is preferred to "defer biopsy" in HR cancer state. Using this notation, we now give a sufficient condition for which there exists a control-limit policy in this context.

Corollary 1. *Denote T as the end of time horizon. Suppose $\eta + \gamma\theta > \theta$, if*

$$(\gamma n - 1)\theta > (\eta + \gamma\theta)\frac{\gamma - \gamma^{n-1}}{1 - \gamma}$$

then there exists an optimal policy at time $T - n$ that is a control-limit type policy for $n = 1, ..., T-1$.

Proof. It is easy to calculate that at $t = T - n$, the smallest possible value at $b = 1$ of choosing "conduct biopsy" is $\eta + \gamma\theta + \gamma n\theta$, where the biopsy result shows not upgrading and "defer biopsy" will be chosen for all future times; the largest possible value at $b = 1$ of choosing "defer biopsy" is $\theta + \eta + \gamma\theta + \frac{1-\gamma^n}{1-\gamma}$, where "conduct biopsy" will be chosen for all future times with the observations all being not upgrading. If

$$\eta + \gamma\theta + \gamma n\theta > \theta + \eta + \gamma\theta + \frac{1-\gamma^n}{1-\gamma},$$

i.e., $(\gamma n - 1)\theta > (\eta + \gamma\theta)\frac{\gamma - \gamma^{n-1}}{1-\gamma}$, then at $b = 1$, the two sets of α-vectors corresponding to two actions are separable. By Lemma 2, we have the optimal policy for the two-state AS-POMDP model is a control-limit type policy. □

The existence of control-limit type policies in practical applications such as ours is a desirable feature since such policies conform well with the intuition of decision-makers. The sufficient condition in the above corollary holds for cases in which γ or θ approaches zero, for instance; however, we show that the existence of a control-limit type policy can be extended more broadly to a special (but not unrealistic) case of our model.

Proposition 3. *For the two-state AS-POMDP model, if decisions are made independent of the PSA test, the optimal policy is a control-limit type policy.*

Proposition 3 aligns well with clinical evidence that the PSA test is associated with high false-positive and false-negative errors and thus plays a limited role in making decisions about when to conduct routine biopsies. The proof of Proposition 3 is shown in the Appendix.

5.2. Static vs. Dynamic Policy

Our computational results in the next section show that although the optimal (dynamic) biopsy policies from the AS-POMDP model dominate the current (static) biopsy guidelines in the published literature, the difference is relatively small. Therefore, we conclude this section with some analysis to explain this by showing that eliminating the PSA test from the model makes it optimal to make biopsy decisions a priori without the need for dynamic decision making. In other words, the schedule of biopsies can be set at the time of diagnosis. Combining this with the fact that PSA is associated with high false-positive and false-negative rates and thus provides limited information for belief updating over time, suggests that the weakness of the PSA test limits the benefits of dynamic changes to the sequential decision to biopsy over time.

Theorem 3. *Consider a threshold-based biopsy policy for AS. If PSA test results are not used in cancer progression belief updates, then the threshold-based policy is equivalent to a static policy, in which the biopsy schedule is pre-determined at the time of diagnosis.*

Theorem 3 provides motivation for why the difference between dynamic and static policies is small, i.e., because the predictive value of the PSA test is weak. The proof of Theorem 3 is in

Center	misclassification error at diagnosis: b_1	Annual Cancer Progression rate: p	Biopsy Sensitivity: $(1-\gamma)$
JH	0.0583	0.0691	0.7184
UCSF	0.0809	0.1217	0.7431
U of T	0.0774	0.1016	0.7949
PRIAS	0.0653	0.0841	0.7614

Table 2: AS-POMDP model parameters in four study centers. Abbreviations: JH, Johns-Hopkins; UCSF, University of California-San Francisco; U of T, University of Toronto; PRIAS, Prostate Cancer Research International Active Surveillance.

the Appendix. Note that tests with better predictive performance than the PSA test, such as new molecular biomarker tests that are being developed (Barnett et al., 2018b), could lead to more significant benefits of dynamic over static policies. We revisit this in Section 6 with numerical experiments.

6. Results

In this section, we discuss the results of the AS-POMDP model for prostate cancer AS. We start by describing the model parameters. Next, we present the results for the near-optimal value function and risk thresholds for the optimal biopsy policy given by the proposed AS-POMDP model using the algorithms in Section 4.2. These results also demonstrate the utility of the approximation methods we proposed. We also illustrate how the AS-POMDP model-based policy changes with respect to the reward parameters to understand how decisions might vary depending on patient preferences. Finally, we compare the near-optimal approximate policies with published guidelines.

6.1. Model Parameters

Tables 2 and 3 provide the As-POMDP model parameters for different centers that are computed using hidden Markov models obtained in a previous study by Li et al. (2020). The PSA distributions were estimated by a mixture of two Gaussian distributions. In our AS-POMDP formulation, we discretized these continuous distributions using commonly used clinical thresholds, as shown in Table 3.

6.2. Optimal Biopsy policy Solved by AS-POMDP Model

The optimal policies of the AS-POMDP model vary across different centers, and reward parameters, which in turn depends on the decision-maker's preference. In our initial experiments, we set $\theta = \eta = -0.5$, which weighs the two criteria, i.e., expected delay in detection of high-risk cancer and expected number of biopsies, equally, and we evaluate the variation in policies across centers.

Figure 2 shows the approximate optimal value functions obtained by the method described in Section 4.2.1, for all four study centers assuming a patient at age 50. Here B_t is chosen to be

Center	Probability Mass Function of PSA (ng/mL): q			
	Cancer State	$I_1 = [0, 4]$	$I_2 = (4, 10]$	$I_3 = (10, \infty)$
JH	LR Cancer	0.3552	0.4311	0.2137
	HR Cancer	0.2868	0.4706	0.2426
UCSF	LR Cancer	0.0768	0.5680	0.3552
	HR Cancer	0.0678	0.5736	0.3586
U of T	LR Cancer	0.4573	0.3422	0.2005
	HR Cancer	0.3312	0.2368	0.4320
PRIAS	LR Cancer	0.1361	0.5357	0.3282
	HR Cancer	0.1094	0.5501	0.3405

Table 3: The probability mass functions of PSA in four study centers. Abbreviations: JH, Johns-Hopkins; UCSF, University of California-San Francisco; U of T, University of Toronto; PRIAS, Prostate Cancer Research International Active Surveillance; LR, low-risk; HR, high-risk.

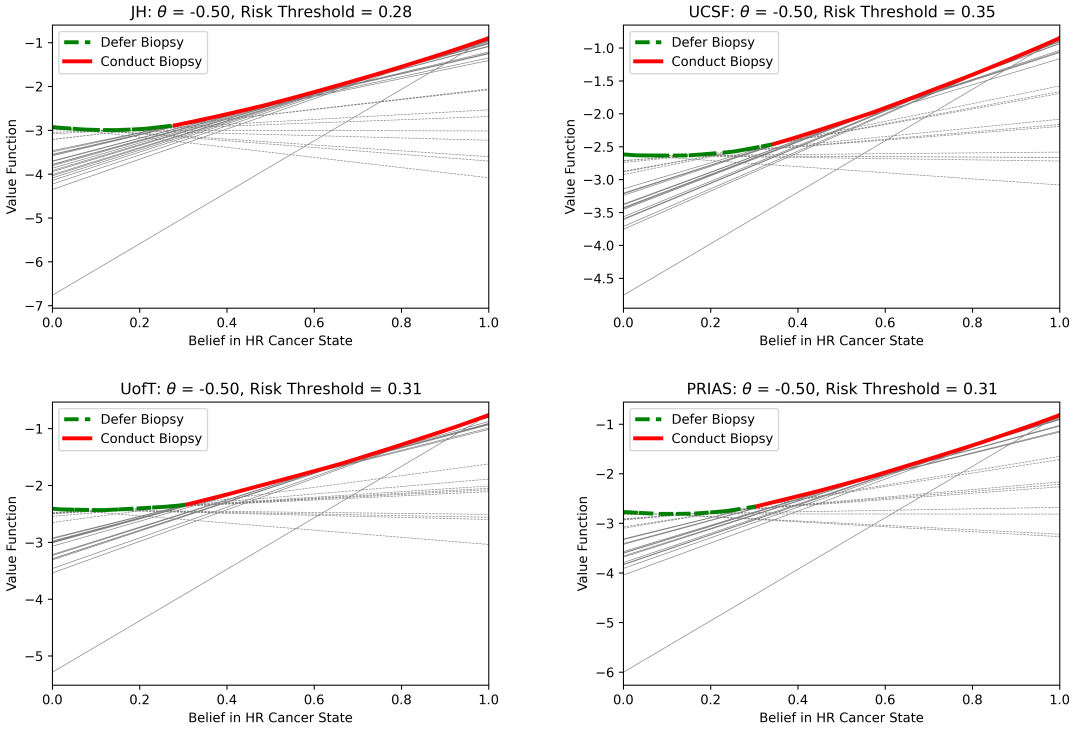

Figure 2: The (approximate) optimal value functions for a patient at age 50 in four different study centers when $\theta = -0.5$. All non-dominated hyperplanes, and their supremums are shown in the figure. The belief threshold for conducting a biopsy is indicated in the legend in each plot.

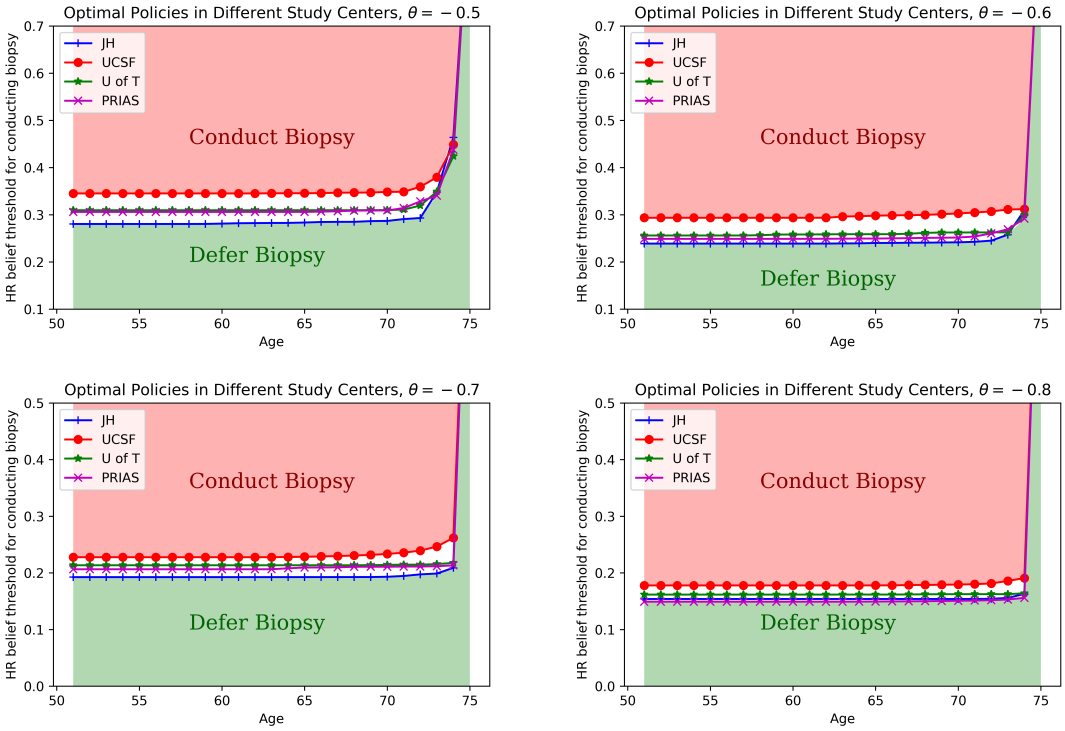

Figure 3: The (approximate) optimal belief thresholds for conducting biopsy in different AS studies when $\theta = -0.5, -0.6, -0.7, -0.8$

$B_t = \{0, \frac{1}{30}, \frac{2}{30}, ..., \frac{29}{30}, 1\}$ with $|B_t| = 31$ for every $t \leq T$. As anticipated, the AS-POMDP model-based policies are all control-limit type policies. The risk threshold for triggering biopsy was the highest in the model generated from UCSF medical center data, and lowest in the model generated from JH hospital data, which is consistent with the difference in the annual cancer progression rates at those centers, and which in turn depends on the study admission criteria (JH study patients had more strict criteria for entry compared to UCSF patients). We further use Figure 3 to illustrate how AS-POMDP model-based policies differ across AS studies. For each risk-based policy, the range of risk threshold is relatively small.

As discussed in Section 3, our AS-POMDP model trades off the two competing criteria (delay in detection vs. harm from biopsies) based on the reward parameter θ. Therefore, Figure 3 also shows how the optimal biopsy policies vary with respect to θ, which is the reward weights of the two criteria depending on individual patient's preference. Again, the closer the θ is to -1, the more the decision-maker weighs on the cost of delay in detection. As we change the value of θ parameter in the proposed AS-POMDP model, we observed that the optimal biopsy policy at each decision epoch is always a control limit type policy as discussed in relation to Proposition 3. Figure 3 also shows that the variation across models derived from the different AS studies decreases as theta decreases, i.e., as the weight on number of biopsies decreases. Moreover, the threshold for biopsy is consistently below 0.4 for all ages prior to 73.

Centers	$\|\|(\bar{V}-\hat{V})/\bar{V}\|\|_\infty \times 100\%$ at age 50 for different θ				
	-0.5	-0.6	-0.7	-0.8	-0.9
JH	0.27%	0.21%	0.15%	0.55%	0.28%
UCSF	0.15%	0.09%	0.08%	0.10%	$< 0.01\%$
U of T	0.19%	0.25%	0.16%	0.10%	$< 0.01\%$
PRIAS	0.18%	0.17%	0.25%	0.09%	0.01%

Table 4: The relative difference between $\bar{V}$ and $\hat{V}$ at age 50 for different θ in four AS studies.

6.3. Accuracy of Approximate Policies

To demonstrate that the approximated policies are very close to optimal, Table 4 provides the supremum norm of the difference between the lower and upper bounds of the optimal value function solved by the approximation methods in Section 4.2 using the uniform grid B_t with $|B_t| = 31$ for every $t \leq T$. As we can see from Table 4, the maximum relative error across all experiments is less than 0.55% of the value function, indicating the approximate policies are sufficiently accurate to be trusted. In terms of the running time, each experiment in Table 4 is completed within 30 seconds (compared with more than 24 hours for an exact solution) using an Intel Core i7 2.6 GHz processor with 16 GB RAM. Thus, the approximations enable the potential real-time implementation of the AS-POMDP model for shared patient/physician decision-making in clinical settings.

6.4. Implementation of Model-based Biopsy Policy in Practice

Before comparing different biopsy policies, we explain how the model-based policy can be used in practice to support decision-making in prostate cancer AS. In each study center, for each patient newly diagnosed with LR prostate cancer and admitted to AS, his initial belief of being in HR cancer state is estimated by the misclassification error at diagnosis in Table 2. Subsequently, at each annual time period, the patient first receives a PSA test and the belief is updated using Equation 1. Next, the decision-maker decides whether to conduct or defer biopsy using an instance of the model based on the choice of the reward parameter θ that aligns with the patient's preferences, and the corresponding optimal HR belief threshold for triggering a biopsy base on the AS-POMDP model. If a biopsy is conducted, as shown in Figure 1, the patient will stay on AS if the result shows no biopsy upgrading and his age is less than 75 (the clinically recommended stopping time). The belief of HR cancer state is then updated again based on the annual cancer progression rate and biopsy sensitivity given by Table 2 using Equation 1; otherwise if the biopsy is deferred, then the HR cancer belief is updated only based on the annual cancer progression rate. Lastly, the patient will continue to the next time period, and follow the same steps as in the last time period until a biopsy upgrading is observed or age 75. We acknowledge that in practice, the decision of whether to conduct biopsy or not is often more nuanced, and requires a shared decision-making approach between the patient and physician. But our model-based biopsy policy can be used as a data-driven decision support tool to guide these decisions.

6.5. Comparison of Model-based Biopsy Policies vs. Current guidelines

Now, we compare the policies from solving the AS-POMDP model with published guidelines. The published guidelines include annual biopsy (JH guideline), biopsy every two years after diagnosis (UCSF guideline), biopsy every three years after diagnosis (PRIAS guideline, which is also implemented in the U of T study). We evaluate each policy for a simulated cohort of patients diagnosed with LR cancer who initiated AS at age 50. We first sample the initial cancer state at the starting time according to the misclassification error at diagnosis given in Table 2. Then, the patients will follow the process described in Figure 1, where at each decision epoch, the test action is given by the selected biopsy policy, the test results are sampled according to the observation probabilities, and the state transition is sampled according to the state transition probability. If a Gleason score upgrading is observed, the patient will leave AS immediately; otherwise, he continues to the next decision epoch, until age 75 when AS stops.

The number of hypothetical patients for the simulation is 10,000 for each study center and each biopsy policy. With the simulated true cancer states and biopsy results for all patients at all decision epochs, the expected number of biopsies performed while on AS is calculated as the average number of biopsies performed from initiating AS (age 50) to leaving AS (age 75 or a Gleason score upgrading), while the expected delay in time to detection of non-favorable risk cancer is calculated as the average difference between the time of the first sampled HR cancer state and the time of a Gleason score upgrading is observed for all patients.

Figure 4 illustrates the simulation results for different biopsy policies in four study centers. As we can see from Figure 4, in each center, for the optimal biopsy policies given by the AS-POMDP model, as the value of $|\theta|$ gets larger, the biopsy policy will result in a greater number of expected biopsies and fewer years to the detection of cancer progression. Also, the optimal biopsy policies given by the AS-POMDP model are Pareto optimal compared with the static biopsy guidelines, i.e., they reduce the number of biopsies performed without increasing years in late detection to cancer progression.

6.6. Using MRI for Active Surveillance

Since the PSA test has high false positive and negative rates, as previously noted, we do not observe a huge improvement in the policy given by the AS-POMDP model over current biopsy guidelines for each patient in Figure 4. Nevertheless, it is possible that more accurate bio-markers could lead to more significant improvement of the AS-POMDP model-based policies over current biopsy guidelines. One such approach to improving predictive performance that is receiving significant attention is MRI. Barnett et al. (2018b) showed the cost-effectiveness of magnetic resonance imaging (MRI) for early detection of prostate cancer. Motivated by their study, we conducted experiments using MRI as an alternative to the PSA test in the AS-POMDP model to show the potential benefit of the model-based policy. For MRI model parameters, we used the result from Grey et al. (2015), which estimated the sensitivity and specificity of MRI (using the prostate imaging reporting and data system score threshold of ≥ 4) to be 78.9% and 78.9%. Figure 5 shows

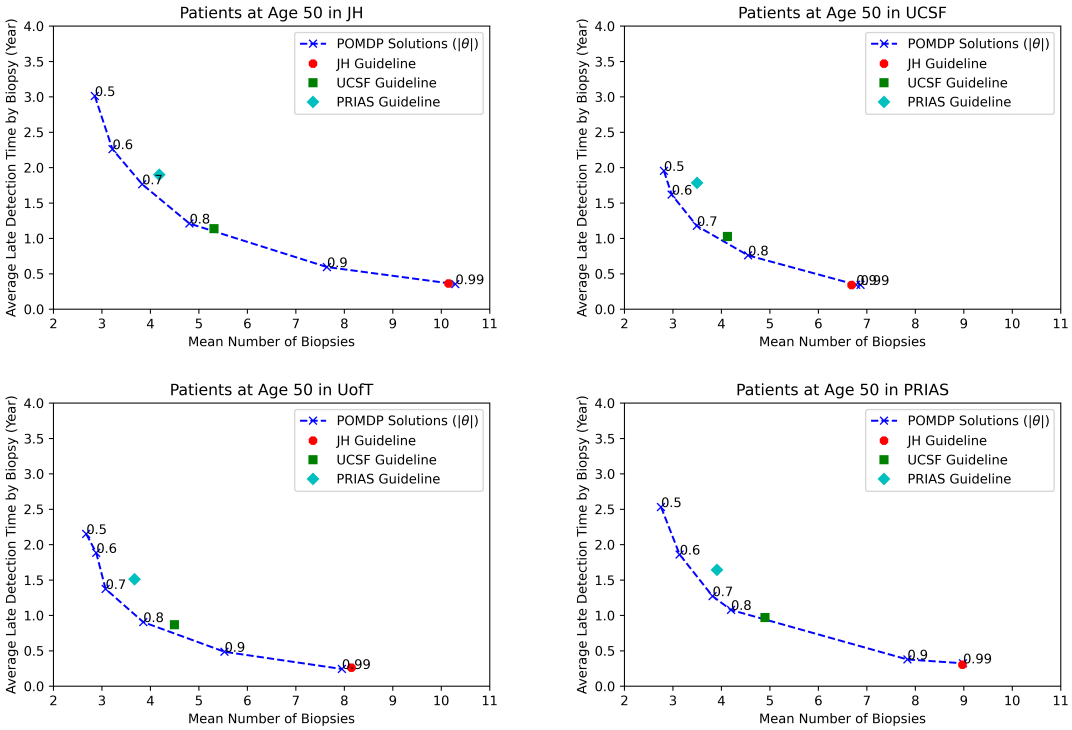

Figure 4: The comparison between policies given by the AS-POMDP model and current biopsy guidelines in different AS studies.

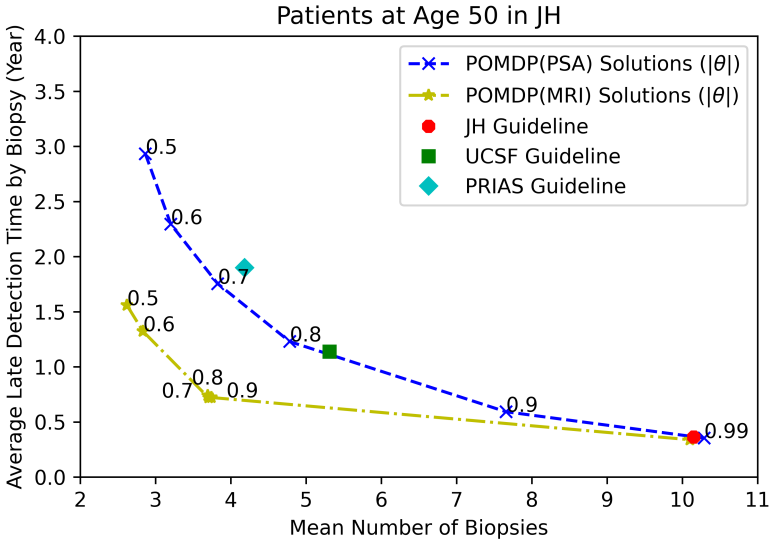

Figure 5: The comparison between policies given by two AS-POMDP (PSA and MRI) models and current biopsy guidelines in the JH center.

the comparison among the policy given by the AS-POMDP model with either PSA test or MRI, and current biopsy guidelines for patients in the JH center. As we can see in Figure 5, as MRI is much more accurate than the PSA test, the benefit of the policies given by the AS-POMDP(MRI) model is more significant than it given by the AS-POMDP(PSA) model. Unfortunately, the study of Grey et al. (2015) was conducted on a different group of patients in the U.K., with a limited size of study population ($n = 201$), so that the result in Figure 5 is from a hypothetical experiment. We are looking forward to implement the MRI in the AS-POMDP when more MRI data and studies become available.

6.7. Evaluating Implied Weights for Late Detection of Cancer Progression and Biopsy Burden

To understand how alternative policies trade-off between late detection to cancer progression and biopsy burden, we apply a simple *inverse optimization* (Ng et al., 2000) to estimate the reward function implied by each published biopsy guideline. Specifically, for a given biopsy guideline, denote $\pi = (\pi_1, ..., \pi_T)$ as the biopsy policy specified by the guideline. Since π is a static policy, then π_t is a constant action w.r.t. the belief state (either to defer biopsy or conduct biopsy) for all $t = 1, ..., T$. Now, denote $\bar{\pi}^t = (\bar{\pi}^t_1, ..., \bar{\pi}^t_T)$ as another static biopsy policy where

$$\pi_t \neq \bar{\pi}^t_t, \text{ and } \pi_k = \bar{\pi}^t_k, \ \forall k \neq t.$$

Further, define R_t as the set of reward functions such that $\bar{\pi}^t$ is dominated by π:

$$R_t := \{r : V^{\pi}_1(b_1) \geq V^{\bar{\pi}^t}_1(b_1)\}, \ \forall t = 1, ..., T,$$

where b_1 is the initial belief state. Notice that the reward function r is a function of θ, and the range of θ where the biopsy guideline π is the optimal static biopsy policy is given by

$$\theta \in R_1 \cap ... \cap R_T.$$

Table 5 shows the estimated range of θ implied by each guideline if applied to each center. As we can see from Table 5, all four study centers imply that avoiding delays in detecting high-risk prostate cancer is more important than avoiding biopsies; however, the relative weights vary significantly among the guidelines, which depend on the cancer progression rate and biopsy sensitivity in different study centers. Nevertheless, as some patients are highly averse to biopsies (Klotz, 2013), our study provides a solution to deciding the frequency of biopsy and a reference for the trade-off against the late detection to a cancer progression.

7. Conclusions

In this paper, we proposed a finite-horizon two-state POMDP (AS-POMDP) model to optimize the biopsy policy in prostate cancer AS, where the objective is to minimize the number of biopsies and the delay in detection of high-risk cancer. Our study considered two kinds of parameter ambiguity: 1) heterogeneous transition and observation probabilities in different patient cohorts, and

Center	Range of θ implied by the biopsy guideline		
	JH guideline	UCSF guideline	PRIAS guideline
JH	$[-1, -0.93]$	$[-0.84, -0.83]$	$[-0.72, -0.71]$
UCSF	$[-1, -0.89]$	$[-0.75, -0.74]$	$[-0.68, -0.67]$
U of T	$[-1, -0.91]$	$[-0.83, -0.82]$	$[-0.78, -0.77]$
PRIAS	$[-1, -0.92]$	$[-0.83, -0.82]$	$[-0.71, -0.70]$

Table 5: Estimates of the range of θ implied by each published biopsy guideline in different AS study centers.

2) variation in decision-maker's preferences as represented by reward functions. To evaluate alternative policies resulting from different parameters, it was necessary to solve many instances of the AS-POMDP model. To enable this, we introduced two fast approximation methods that are able to find the lower and upper bounds of the optimal value function of the AS-POMDP model. We compared the gap between the lower and upper bounds to show that our results were accurate enough for decision-making. Further, We discussed some structural properties of the AS-POMDP model that provide insight into the AS-POMDP model-based policies. We also discussed an explanation for why the dynamic biopsy policies given by the AS-POMDP model are similar to static policies recommended in the current biopsy guidelines, and we used inverse optimization to approximate how each guideline weighs biopsy burden versus late detection of cancer progression.

In the computational result, we first presented the value functions and biopsy policies given by the AS-POMDP model in four different prostate cancer AS studies, if weighted equally on the burden of one biopsy and the penalty of one-year late detection to cancer progression. We observed that the optimal value function is not always monotone in the belief state. This is because the objective of the AS-POMDP model is to investigate rather than improve patients' cancer state, and patients may leave the system without any future cost if detected as high-risk cancer. Such models can be more straightforward for studies of medical testing, and more accurate, especially when other metrics such as QALYs are hard to estimate and too obscure for decision-making. Although the optimal value function is not monotone, we observe that the biopsy policies given by the optimal value function were monotone in the belief in high-risk cancer state, i.e., it would trigger a biopsy as long as the belief in the high-risk cancer state reached a threshold. The threshold of the optimal biopsy policy is dependent on the model parameters, which include cancer progression rate and biopsy sensitivity. In general, models with a higher cancer progression rate or lower biopsy sensitivity will give a lower belief threshold for conducting biopsy.

We then changed the reward weights in the reward function to see how does the model-based biopsy policy depends on the decision-maker's preference on biopsy burden and late detection time in each study center. We found that the more heavily the decision-maker weighs the late detection of cancer progression (the larger θ), the lower the belief threshold for triggering a biopsy in the optimal biopsy policy.

Finally, we compared the performance of the optimal biopsy policies given by the AS-POMDP

model and current biopsy guidelines in four AS study centers by a simulation study. The model-based biopsy policies were all Pareto optimal. The policies based on published guidelines were close to the efficient frontier. We also ran a hypothetical experiment using MRI in the AS-POMDP model, which showed the potential value of the AS-POMDP model with more accurate bio-markers than PSA. Lastly, we used an inverse optimization approach to estimate the reward weights implied by the current biopsy guidelines.

Besides the novelty of the application, our work also contributes to the POMDP literature. First, we introduced two fast approximation methods to quickly find the lower and upper bounds of the optimal value function of a finite-horizon POMDP model at each decision epoch. In particular, we showed that the best upper bound of the optimal value function at any belief point could be solved easily in a two-state model, without solving a large linear program as discussed in previous studies. We also provided the worst-case error bounds of the proposed approximation methods. Second, we showed that in extreme cases, the optimal biopsy policy given by the AS-POMDP model is a control-limit type policy, even if the optimal value function is not monotone in the belief state, which differs from all previous studies of the control-limit type policy. We discussed some intermediate results for the sufficient and necessary condition for the existence of a control-limit type policy in the POMDP model. We leave the statement for general cases as a conjecture in future research. Third, we showed that in the proposed AS-POMDP model, if the PSA test is not involved, then the optimal dynamic policy given by the model is equivalent to a static policy, in which the timing of conducting biopsy can be pre-determined. Further, we applied inverse optimization to approximate the value function implied by the current biopsy guidelines, which helped us understand how does each biopsy guideline weigh on late detection of cancer progression and biopsy burden.

There are also some limitations of our work, which could lead to opportunities for future research. First, we used a two-state POMDP model to approximate the stochastic system of prostate cancer AS, and only considered the information from PSA test and biopsy. There might be other covariates in prostate cancer AS such as prostate volume, PSA doubling time, and the results of MRI scans that could be used to understand the underlying cancer state, but were not considered in this study. We look forward to improving our model by including these factors when more data becomes available. Second, the model parameters of the transition and observation probabilities are assumed to be stationary, i.e., independent of time, which may not be accurate in reality. However, incorporating time-dependent factors would require the estimates of the model parameters in pre-studies, and more computational effort to solve the model. Third, our results of the fast approximation method for finding the upper bound of the optimal value function, and the sufficient and necessary condition for the existence of a control-limit type policy only work in a two-state POMDP model. The generalization of these results to general POMDP models may not be trivial and is left for future studies. Although the focus of this article is on prostate cancer AS, our model formulation is flexible and could be applied to other medical decision-making problems in chronic disease management.

Acknowledgements

This material is based upon work supported in part by the National Science Foundation through Grant Number CMMI 0844511. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. This work was also supported by the Movember Foundation. The funders did not play any role in the study design, collection, analysis or interpretation of data, or in the drafting of this paper.

References

Albright, S. C. (1979). Structural results for partially observable markov decision processes. *Operations Research*, *27*, 1041–1053.

Anandadas, C. N., Clarke, N. W., Davidson, S. E., O'Reilly, P. H., Logue, J. P., Gilmore, L., Swindell, R., Brough, R. J., Wemyss-Holden, G. D., Lau, M. W. et al. (2011). Early prostate cancer–which treatment do men prefer and why? *BJU international*, *107*, 1762–1768.

Åström, K. J. (1965). Optimal control of markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, *10*, 174–205.

Ayer, T., Alagoz, O., & Stout, N. K. (2012). Or forum—a pomdp approach to personalize mammography screening decisions. *Operations Research*, *60*, 1019–1034.

Ayer, T., Alagoz, O., Stout, N. K., & Burnside, E. S. (2016). Heterogeneity in women's adherence and its role in optimal breast cancer screening policies. *Management Science*, *62*, 1339–1362.

Barnett, C. L., Auffenberg, G. B., Cheng, Z., Yang, F., Wang, J., Wei, J. T., Miller, D. C., Montie, J. E., Mamawala, M., & Denton, B. T. (2018a). Optimizing active surveillance strategies to balance the competing goals of early detection of grade progression and minimizing harm from biopsies. *Cancer*, *124*, 698–705.

Barnett, C. L., Davenport, M. S., Montgomery, J. S., Wei, J. T., Montie, J. E., & Denton, B. T. (2018b). Cost-effectiveness of magnetic resonance imaging and targeted fusion biopsy for early detection of prostate cancer. *BJU international*, *122*, 50–58.

Bastian, P. J., Carter, B. H., Bjartell, A., Seitz, M., Stanislaus, P., Montorsi, F., Stief, C. G., & Schröder, F. (2009). Insignificant prostate cancer and active surveillance: from definition to clinical implications. *European urology*, *55*, 1321–1332.

Bul, M., Zhu, X., Valdagni, R., Pickles, T., Kakehi, Y., Rannikko, A., Bjartell, A., Van Der Schoot, D. K., Cornel, E. B., Conti, G. N. et al. (2013). Active surveillance for low-risk prostate cancer worldwide: the prias study. *European urology*, *63*, 597–603.

Cassandra, A., Littman, M. L., & Zhang, N. L. (1997). Incremental pruning: A simple, fast, exact method for partially observable markov decision processes. In *In Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence* (pp. 54–61). Morgan Kaufmann Publishers.

Cassandra, A. R. (1998). A survey of pomdp applications. In *Working notes of AAAI 1998 fall symposium on planning with partially observable Markov decision processes*. volume 1724.

Cassandra, A. R., Kaelbling, L. P., & Kurien, J. A. (1996). Acting under uncertainty: Discrete bayesian models for mobile-robot navigation. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS'96* (pp. 963–972). IEEE volume 2.

Coley, R. Y., Fisher, A. J., Mamawala, M., Carter, H. B., Pienta, K. J., & Zeger, S. L. (2017). A bayesian hierarchical model for prediction of latent health states from multiple data sources with application to active surveillance of prostate cancer. *Biometrics*, *73*, 625–634.

Dall'Era, M. A., Cooperberg, M. R., Chan, J. M., Davies, B. J., Albertsen, P. C., Klotz, L. H., Warlick, C. A., Holmberg, L., Bailey Jr, D. E., Wallace, M. E. et al. (2008). Active surveillance for early-stage prostate cancer: review of the current literature. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, *112*, 1650–1659.

Dall'Era, M. A., Albertsen, P. C., Bangma, C., Carroll, P. R., Carter, H. B., Cooperberg, M. R., Freedland, S. J., Klotz, L. H., Parker, C., & Soloway, M. S. (2012). Active surveillance for prostate cancer: a systematic review of the literature. *European urology*, *62*, 976–983.

Drake, A. W. (1962). *Observation of a Markov process through a noisy channel*. Ph.D. thesis Massachusetts Institute of Technology.

Epstein, J. I., Feng, Z., Trock, B. J., & Pierorazio, P. M. (2012). Upgrading and downgrading of prostate cancer from biopsy to radical prostatectomy: incidence and predictive factors using the modified gleason grading system and factoring in tertiary grades. *European urology*, *61*, 1019–1024.

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

Grey, A. D., Chana, M. S., Popert, R., Wolfe, K., Liyanage, S. H., & Acher, P. L. (2015). Diagnostic accuracy of magnetic resonance imaging (mri) prostate imaging reporting and data system (pi-rads) scoring in a transperineal prostate biopsy setting. *BJU international*, *115*, 728–735.

Hauskrecht, M. (2000). Value-function approximations for partially observable markov decision processes. *Journal of artificial intelligence research*, *13*, 33–94.

Hoffman, R. M. (2011). Screening for prostate cancer. *New England Journal of Medicine*, *365*, 2013–2019.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial intelligence*, *101*, 99–134.

Klotz, L. (2010). Active surveillance for prostate cancer: a review. *Current urology reports*, *11*, 165–171.

Klotz, L. (2013). Prostate cancer overdiagnosis and overtreatment. *Current Opinion in Endocrinology, Diabetes and Obesity*, *20*, 204–209.

Klotz, L., Zhang, L., Lam, A., Nam, R., Mamedov, A., & Loblaw, A. (2009). Clinical results of long-term follow-up of a large, active surveillance cohort with localized prostate cancer. *Journal of Clinical Oncology*, *28*, 126–131.

Li, W., Denton, B. T., Nieboer, D., Carroll, P. R., Roobol, M. J., Morgan, T. M., & consortium, M. F. G. A. P. P. C. A. S. G. (2020). Comparison of biopsy under-sampling and annual progression using hidden markov models to learn from prostate cancer active surveillance studies. *Cancer Medicine*, .

Littman, M. L., Cassandra, A. R., & Kaelbling, L. P. (1995). *Efficient dynamic-programming updates in partially observable Markov decision processes*. Technical Report Brown University.

Lovejoy, W. S. (1987). Some monotonicity results for partially observed markov decision processes. *Operations Research*, *35*, 736–743.

Lovejoy, W. S. (1991). A survey of algorithmic methods for partially observed markov decision processes. *Annals of Operations Research*, *28*, 47–65.

Miehling, E., & Teneketzis, D. (2020). Monotonicity properties for two-action partially observable markov decision processes on partially ordered spaces. *European Journal of Operational Research*, *282*, 936–944.

Ng, A. Y., Russell, S. J. et al. (2000). Algorithms for inverse reinforcement learning. In *Icml* (p. 2). volume 1.

Otten, M., Timmer, J., & Witteveen, A. (2020). Stratified breast cancer follow-up using a continuous state partially observable markov decision process. *European journal of operational research*, *281*, 464–474.

Pineau, J., Gordon, G., Thrun, S. et al. (2003). Point-based value iteration: An anytime algorithm for pomdps. In *IJCAI* (pp. 1025–1032). volume 3.

Ross, S. M. (1971). Quality control under markovian deterioration. *Management Science*, *17*, 587–596.

Sandıkçı, B., Maillart, L. M., Schaefer, A. J., & Roberts, M. S. (2013). Alleviating the patient's price of privacy through a partially observable waiting list. *Management Science*, *59*, 1836–1854.

Shani, G., Pineau, J., & Kaplow, R. (2013). A survey of point-based pomdp solvers. *Autonomous Agents and Multi-Agent Systems*, *27*, 1–51.

Simmons Ivy, J., Black Nembhard, H., & Baran, K. (2009). Quantifying the impact of variability and noise on patient outcomes in breast cancer decision making. *Quality Engineering*, *21*, 319–334.

Smallwood, R. D., & Sondik, E. J. (1973). The optimal control of partially observable markov processes over a finite horizon. *Operations research*, *21*, 1071–1088.

Spaan, M. T., & Vlassis, N. (2005). Perseus: Randomized point-based value iteration for pomdps. *Journal of artificial intelligence research*, *24*, 195–220.

Thomsen, F. B., Brasso, K., Klotz, L. H., Røder, M. A., Berg, K. D., & Iversen, P. (2014). Active surveillance for clinically localized prostate cancer—-a systematic review. *Journal of surgical oncology*, *109*, 830–835.

Tosoian, J. J., Trock, B. J., Landis, P., Feng, Z., Epstein, J. I., Partin, A. W., Walsh, P. C., & Carter, H. B. (2011). Active surveillance program for prostate cancer: an update of the johns hopkins experience. *J Clin Oncol*, *29*, 2185–2190.

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

White, C. C. (1979). Optimal control-limit strategies for a partially observed replacement problem. *International Journal of Systems Science*, *10*, 321–332.

White, C. C. (1991). A survey of solution techniques for the partially observed markov decision process. *Annals of Operations Research*, *32*, 215–230.

Zhang, J., Denton, B. T., Balasubramanian, H., Shah, N. D., & Inman, B. A. (2012a). Optimization of prostate biopsy referral decisions. *Manufacturing & Service Operations Management*, *14*, 529–547.

Zhang, J., Denton, B. T., Balasubramanian, H., Shah, N. D., & Inman, B. A. (2012b). Optimization of psa screening policies: a comparison of the patient and societal perspectives. *Medical Decision Making*, *32*, 337–349.

Zhang, N. L., & Liu, W. (1996). *Planning in stochastic domains: Problem characteristics and approximation*. Technical Report Technical Report HKUST-CS96-31, Hong Kong University of Science and Technology.

Appendix

Proof of Theorem 1

Proof. First,

$$\begin{aligned}||V_t - \hat{V}_t||_\infty &= ||HV_{t+1} - L^B\hat{V}_{t+1}||_\infty \\ &= ||HV_{t+1} - H\hat{V}_{t+1} + H\hat{V}_{t+1} - L^B\hat{V}_{t+1}||_\infty \\ &\leq ||HV_{t+1} - H\hat{V}_{t+1}||_\infty + ||H\hat{V}_{t+1} - L^B\hat{V}_{t+1}||_\infty\end{aligned}$$

For the first term, $||HV_{t+1} - H\hat{V}_{t+1}||_\infty \leq ||V_{t+1} - \hat{V}_{t+1}||_\infty$. For the second term, let $b \in \mathcal{B}$ be the belief point where the point-based value approximation has the biggest error, and $\tilde{b} \in B$ be the closest sampled belief point to b. Also, let α be the vector that would be the maximal at b, and $\tilde{\alpha}$ be the vector that is maximal at $\tilde{b}$, then $\tilde{\alpha} \cdot \tilde{b} \geq \alpha \cdot \tilde{b}$, and

$$\begin{aligned}||HV_{t+1}^B - L^B V_{t+1}^B||_\infty &\leq \alpha \cdot b - \tilde{\alpha} \cdot b \\ &= \alpha \cdot b - \tilde{\alpha} \cdot b + (\alpha \cdot \tilde{b} - \alpha \cdot \tilde{b}) \\ &\leq \alpha \cdot b - \tilde{\alpha} \cdot b + (\tilde{\alpha} \cdot \tilde{b} - \alpha \cdot \tilde{b}) \\ &= (\alpha - \tilde{\alpha}) \cdot (b - \tilde{b}) \\ &\leq ||\alpha - \tilde{\alpha}||_\infty ||b - \tilde{b}||_1\end{aligned}$$

where the last step is by the Holder's inequality. Now, since each α-vector represents the cumulative reward from the current time until the end of time horizon followed by a policy specifying the choices of future actions for all possible observation sequences, then

$$||\alpha - \tilde{\alpha}||_\infty \leq (T-t)(r_{\max} - r_{\min}),$$

and

$$||HV_{t+1}^B - L^B V_{t+1}^B||_\infty \leq (T-t)(r_{\max} - r_{\min})\delta.$$

Repeat the steps above, we have

$$\begin{aligned}||V_t - \hat{V}_t||_\infty &\leq ||V_{t+1} - \hat{V}_{t+1}||_\infty + (T-t)(r_{\max} - r_{\min})\delta \\ &\leq |V_{t+2} - \hat{V}_{t+21}||_\infty + [(T-t) + (T-(t+1))](r_{\max} - r_{\min})\delta \\ &\leq ... \\ &\leq \frac{(T-t)(T-t+1)}{2}||r_{\max} - r_{\min}||_\infty \delta.\end{aligned}$$

□

Proof of Theorem 2

Proof. The proof is similar to the proof of Theorem 1. First,

$$\begin{aligned}||V_t - \bar{V}_t||_\infty &= ||HV_{t+1} - U^B\bar{V}_{t+1}||_\infty \\ &= ||HV_{t+1} - H\bar{V}_{t+1} + H\bar{V}_{t+1} - U^B\bar{V}_{t+1}||_\infty \\ &\leq ||HV_{t+1} - H\bar{V}_{t+1}||_\infty + ||H\bar{V}_{t+1} - U^B\bar{V}_{t+1}||_\infty\end{aligned}$$

For the first term, $||HV_{t+1} - H\bar{V}_{t+1}||_\infty \leq ||V_{t+1} - \bar{V}_{t+1}||_\infty$. For the second term, let $b \in \mathcal{B}$ be the belief point where the point-based value approximation has the biggest error, and $\tilde{b} \in B$ be the closest sampled belief point to b. Also, let α be the vector that would be the maximal at b, and $\tilde{\alpha}$ be the vector that is maximal at $\tilde{b}$, then $\tilde{\alpha} \cdot \tilde{b} \geq \alpha \cdot \tilde{b}$, and

$$||HV_{t+1}^B - U^B V_{t+1}^B||_\infty \leq \alpha \cdot b - \tilde{\alpha} \cdot b$$

The rest of the proof is exactly the same as the one for Theorem 1. □

Proof of Proposition 3

Proof. First, it is easy to see that at each decision epoch, the α-vectors of the policy that always chooses "defer biopsy" is a non-dominated α-vector, which achieves a maximum value at $b = 0$ that can be denoted as $x_{t,0}$. Next, we prove by induction that at each decision epoch, all non-dominated α-vectors corresponding to action "defer biopsy" at current time must have their value at $b = 0$ being greater than $x_{t,0} + \eta$. If this statement is true, then the non-dominated α-vectors corresponding to action "defer biopsy" and the non-dominated α-vectors corresponding to action "conduct biopsy" are separable at $b = 0$. By Lemma 2, we have the optimal policy for the two-state AS-POMDP model is a control-limit type policy.

Now, at time T, the α-vectors corresponding to action "defer biopsy" is $(0, \theta)$, and the α-vectors corresponding to action "conduct biopsy" is $(\eta, \eta + \gamma\theta)$.

Assume that at time $t + 1$, all non-dominated α-vectors corresponding to action "defer biopsy" have their value at $b = 0$ being greater that $x_{t+1,0}+\eta$, where $x_{t+1,0}$ is the value at $b = 0$ corresponding to the policy "no biopsy at all". At time t, denote the α-vectors corresponding to policy "no biopsy at all" as $(x_{t,0}, y_{t,0} + \theta)$. Suppose there exists a non-dominated α-vectors corresponding to action "defer biopsy", denoted as $(x_{t,1}, y_{t,1} + \theta)$, such that $x_{t,1} < x_{t,0} + \eta$. We are going to prove that $(x_{t,1}, y_{t,1}+\theta)$ is dominated by others. Consider the α-vectors corresponding to policy "biopsy at time t and no biopsy afterwards", which is $(x_{t,0}+\eta, \gamma y_{t,0}+\eta+\gamma\theta)$. If $(x_{t,1}, y_{t,1}+\theta)$ is not dominated by the maximum of $(x_{t,0}, y_{t,0}+\theta)$ and $(x_{t,0}+\eta, \gamma y_{t,0}+\eta+\gamma\theta)$, then it must be true that the intersect of $(x_{t,0}, y_{t,0} + \theta)$ and $(x_{t,0} + \eta, \gamma y_{t,0} + \eta + \gamma\theta)$, denoted as b_1 is smaller than the intersection of $(x_{t,0} + \eta, \gamma y_{t,0} + \eta + \gamma\theta)$ and $(x_{t,1}, y_{t,1} + \theta)$, denoted as b_2. It is easy to calculate that

$$b_1 = \frac{\eta}{(1-\gamma)(y_{t,0} + \theta)},\ b_2 = \frac{\eta + x_{t,0} - x_{t,1}}{x_{t,0} - x_{t,1} + (y_{t,1} + \theta) - (\gamma\theta + \gamma y_{t,0})}.$$

By backward induction,

$$x_{t,0} = (1-p)x_{t+1,0} + p y_{t+1,0},\ x_{t,1} = (1-p)x_{t+1,1} + p y_{t+1,1}$$

if $x_{t,1} < x_{t,0} + \eta$, since $y_{t+1,0} < y_{t+1,1}$, then $x_{t+1,1} < x_{t+1,0} + \eta$. By assumption, the action at time $t + 1$ corresponding to $(x_{t+1,1}, y_{t+1,1})$ is "conduct biopsy". So, for the α-vector $(x_{t,1}, y_{t,1} + \theta)$, its action at time t and time $t + 1$ are "defer biopsy" and "conduct biopsy". Now, we consider an α-vector at time t, denoted as $(x_{t,2}, y_{t,2})$ whose action at time t and time $t+1$ are "conduct biopsy"

and "defer biopsy", and actions after time $t+1$ are all same as the ones of $(x_{t,1}, y_{t,1}+\theta)$. We can calculate that

$$x_{t,2} = x_{t,1} + p(1-\gamma)(\theta + y_{t+1,2}),\ y_{t,2} = y_{t,1} + (\gamma-1)\theta.$$

Now, we are going to show that $(x_{t,1}, y_{t,1}+\theta)$ is dominated by the maximum of $(x_{t,2}, y_{t,2})$ and $(x_{t,0}+\eta, \gamma y_{t,0}+\eta+\gamma\theta)$. Denote the intersection between $(x_{t,2}, y_{t,2})$ and $(x_{t,0}+\eta, \gamma y_{t,0}+\eta+\gamma\theta)$ as b_3, then

$$b_3 = \frac{x_{t,0}+\eta-x_{t,2}}{x_{t,0}-x_{t,2}+y_{t,2}-\gamma y_{t,0}-\gamma\theta}.$$

Given $b_1 \leq b_2$, it is easy to verify that $b_3 \leq b_2$, which indicated that $(x_{t,1}, y_{t,1}+\theta)$ is dominated by the maximum of $(x_{t,2}, y_{t,2})$ and $(x_{t,0}+\eta, \gamma y_{t,0}+\eta+\gamma\theta)$. In other words, if there exists an α-vector whose optimal action at time t and $t+1$ are "defer biopsy" and "conduct biopsy", then the α-vector should be dominated by another α-vector whose optimal action at time t and $t+1$ are "conduct biopsy" and "defer biopsy". This gives a conflict with the assumption that $(x_{t,1}, y_{t,1}+\theta)$ is non-dominated. As a result, we proved at time t, there is no non-dominated α-vector corresponding to action "defer biopsy" such that its value at $b=0$ is smaller than $x_{t,0}+\eta$.

To sum up, we have proved that at each decision epoch, the non-dominated α-vectors corresponding to action "defer biopsy" and the non-dominated α-vectors corresponding to action "conduct biopsy" are separable at $b=0$. By Lemma 2, we have the optimal policy for the two-state AS-POMDP model is a control-limit type policy. □

Proof of Theorem 3

Proof. We use a straightforward induction argument to show that at each decision epoch, the belief the patient is in the high-risk cancer state can always be pre-calculated whether the biopsy is conducted or deferred at each decision epoch. In the beginning, the patient enters AS with a fixed initial belief of high-risk cancer state b_0. Now, suppose at time t, the patient stays in AS with a fixed belief of high-risk cancer state b_t, then the patient chooses to either choose to do biopsy according to the threshold-based biopsy policy or do nothing. If he chooses to do the biopsy, then he will stay in the AS until the next decision epoch only if the biopsy result is not Gleason score upgrading. So his belief in the high-risk cancer state at time $t+1$ can be calculated by the belief updating formula, which is a fixed value. Otherwise, if he does not perform the biopsy, then his belief of being in the high-risk cancer state at time $t+1$ can be calculated by the state progression formula, which is also fixed. Thus, at each decision epoch t, if the patient does biopsy according to the threshold-based biopsy policy, then his belief in high-risk cancer state is always fixed so that the timing of biopsy is pre-determined. □