ABSTRACT

BUCH, KUANG-HAO. A POMDP Approach for Imputing Rewards to Time-Dependent Breast Cancer Screening Policies. (Under the direction of Julie S. Ivy and James R. Wilson.)

Breast cancer is the most common noncutaneous cancer and the second most common cause of cancer death in the United States. To diagnose breast cancer in the early stage when treatment is most effective, a woman is recommended to adhere to a screening policy (a schedule of regular screening tests) over her lifetime. The variability in the screening policies that have been recommended by different organizations confuse the general public. Hence, this research seeks to estimate the following quantities that are relevant to the use of a given screening policy: (a) the implied cost of premature death from breast cancer expressed in terms of the associated mortality probabilities; (b) the expected total cost of screening mammograms to prevent such premature death expressed in terms of the expected number of mammograms performed over the relevant time horizon; and (c) the trade-off between these costs.

This study develops and implements an inverse algorithm for a partially observable Markov decision process (POMDP)–based model for time-dependent breast cancer screening policies. A time-dependent screening policy refers to a rule for taking action that depends on the timing of the action. POMDPs, widely used in healthcare decision making, are used in this current research because of the indirectly observable nature of a patient's health progression. With a certain level of accuracy, information about the patient's health status is gathered through mammograms. In contrast to regular POMDPs that compute the optimal screening policy for given costs, an inverse POMDP algorithm imputes the costs associated with a given screening policy for which that screening policy is better than all alternative policies.

Degeneracy is an important consideration when developing an inverse algorithm — i.e., the mapping from the policy to the costs is typically not one-to-one. To overcome the degener-

acy issue, we exploit a heuristic method to maximize the difference between the given policy (which is presumed to be "best" with respect to a single overall measure of performance that incorporates all costs and rewards) and the "next best" alternative policy.

We apply our inverse POMDP algorithm to the analysis of several time-dependent screening policies, including the policies currently recommended by the American Cancer Society and the US Preventive Services Task Force. A sensitivity analysis demonstrates the robustness of our imputed costs with respect to uncertainty in key parameters of the POMDP model.

A POMDP Approach for Imputing Rewards to Time-Dependent Breast Cancer Screening Policies

by
Kuang-Hao Buch

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Industrial Engineering

Raleigh, North Carolina

2014

APPROVED BY:

Russel E. King	Reha Uzsoy
Julie S. Ivy Co-chair of Advisory Committee	James R. Wilson Co-chair of Advisory Committee

DEDICATION

To my family.

BIOGRAPHY

Kuang-Hao Buch was born in Yuanlin, Taiwan, R.O.C. In 2002, Kuang-Hao received her Bachelor of Science in Civil Engineering from National Chiao Tung University in Hsinchu, Taiwan. She then received her Master of Science in Industrial and System Engineering from the University of Florida in 2006.

ACKNOWLEDGEMENTS

I would like to thank Dr. Julie S. Ivy and Dr. James R. Wilson for their guidance, support, and encouragement. I especially appreciate Dr. Ivy's support and care for students which allowed me to have the balance between family life and school work, and Dr. Wilson's devotion of his time and effort to develop my research skills. I would also like to acknowledge the support from the Centers for Disease Control and Prevention (CDC NC PERRC grant 1PO1 TP 000296) for providing me with valuable research experience in public health prepareness.

I am grateful to have Dr. Reha Uzsoy as my committee member and mentor for the Preparing the Professoriate program at NC State, whose guidance helped develop my teaching skills. I would like to thank Dr. Russell King for serving as my committee member and for his detailed editing on my dissertation.

I want to thank all my peers in Daniels Hall whose friendship and memories I will cherish for all my life. Also, I want to thank Hui-Ling Liao's family for offering the best support a friend could ever ask for. Finally, I thank my parents-in-law for their understanding and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 The Problem of Screening for Breast Cancer

Cancer is the second most common cause of cancer death in the United States, exceeded only by heart disease [44]. Approximately 580,350 Americans are expected to die of cancer in 2013 [44]. Fortunately, with improvements in cancer treatment and the diagnosis of cancers at an earlier stage, the 5-year relative survival rate for all diagnosed cancers improved from 49% between 1975 and 1977 to 68% between 2002 and 2008 [44]. For most cancers, diagnosing the disease at the earliest possible stage enables the follow-up treatments to deliver the greatest benefit to the affected patients. To diagnose cancer at an earlier stage, screening tests play the most important role. Screening tests for breast cancer provide some of the most compelling evidence of the effectiveness of screening in the early detection and treatment of cancer. The primary objective of this dissertation is to formulate methods for identifying optimal (or near-optimal) breast cancer screening policies and for inferring (imputing) the associated risks of death caused by breast cancer for individuals who adhere to those policies.

Screening is the application of a medical test to detect a potential disease or condition in

an individual who has no known signs or symptoms of that disease or condition [16]. One of the screening test's major objectives is the early detection of disease at a point when treatment is more effective, less expensive, or both [20]. However, as the popular adage says, "there is no such thing as a free lunch," which in the current context (namely, breast cancer screening tests) means there are disadvantages and risks associated with a screening test, such as the probability of an incorrect test result and the radiation risk associated with the test. Hence, when evaluating a screening policy, an important issue is finding the frequency of performing a screening test that best utilizes the benefits of the test while limiting the harms of performing the test.

For the purpose of exploring the benefits of screening, Partially Observable Markov Decision Processes (POMDPs) are widely used in healthcare decision making due to the indirectly observable nature of a patient's health status. The underlying Markov process in a POMDP is used to describe the natural history (progression) of breast cancer in an individual patient. The objective of such a POMDP is to find the optimal screening policy for breast cancer, where a screening policy is a schedule of screening tests over time that specifies the starting age and ending age for screening a patient, the interval between screening tests, etc. However, a policy acquired from a forward POMDP may be hard to implement for the general public because effective implementation of the policy requires the accurate assessment of the belief of each patient's state of health. However, to perform such an assessment requires a well trained medical professional because all types of risk factors associated with the disease are involved in such an assessment. Hence, most of the screening policy recommendations, such as the breast cancer screening policy of the American Cancer Society (ACS) and the cervical cancer screening policy of the US Preventive Services Task Force (USPSTF), depend more on timing than on a patient's health status in absence of any risk factor.

1.2 Objectives of the Research

Different organizations recommend different screening policies, and such conflicting recommendations confuse the general public. For example, the USPSTF recommends an biennial breast cancer screening policy for women between ages 50 and 74 while the ACS recommends an annual screening policy for women from age 40 and beyond. So the proposed research seeks to address the following question: "What does a breast cancer screening policy say about a patient's health with regard to preventing that disease?" From the perspectives of economic analysis, we are naturally led to ask: "Are there some *imputed* rewards or costs for a woman associated with following a particular breast cancer screening policy when a doctor recommends the policy?" For example, the USPSTF recommends biennial breast cancer screening between ages 50 and 74, which may imply that the breast cancer risk of a woman who is younger than 50 or older than 74 is not significant enough to perform a screening test. Therefore, the first goal of this proposed research is to develop an inverse POMDP algorithm for time-dependent breast cancer screening policies, where a time-dependent policy refers to a rule for taking action that depends on the timing of the action before the patient's health status becomes symptomatic. Using this proposed inverse algorithm, we will pursue the second related goal of performing an analysis of the advantages and disadvantages of time-dependent policies given the necessary optimality conditions of time-dependent policies.

In the proposed research, we develop a model for breast cancer applied to the general population and investigate the imputed rewards in terms of a woman's lifetime breast cancer mortality probability — i.e., the probability that for a woman of a given age who is in a given belief health state, her ultimate cause of death will be breast cancer. This model searches for the imputed rewards for a chosen time-dependent screening policy such that the chosen screening policy is better than all other time-dependent policies. To this end, we formulate a POMDP for

breast cancer screening and we apply the inverse algorithm presented in Section 2.5. Owing to the nature of breast cancer, a woman's health status is not directly observable; nevertheless, information about her health status can be gathered through mammography with a certain level of accuracy.

By applying the inverse algorithm to the forward breast cancer POMDP model, we can learn what the reward structure must be so that a particular time-dependent screening policy will yield the largest reduction in a woman's lifetime breast cancer mortality probability. Furthermore, we can use the reward structure to compare different time-dependent policies, allowing the patient to select a time-dependent policy for herself that is most "appropriate" in terms of her belief regarding her breast cancer mortality risk and that corresponds to acceptable levels of the lifetime breast cancer rates for a given range of patient ages and given health states.

1.3 Organization of the Dissertation

The remainder of this research is organized as follows. Chapter 2 focuses on the inverse problem and the algorithm for a time-dependent breast cancer screening policy. In Chapter 2, the concept of the inverse problem is introduced and the difference between a regular problem and an inverse problem is discussed. Then, a literature review of inverse problems and algorithms is provided. The discussion continues by introducing the background on the inverse problem, i.e., a finite-state, discrete-time POMDP is formulated. Following the general POMDP formulation, we present a different POMDP formulation that has been adapted for disease screening. An inverse algorithm for a time-dependent policy is introduced. An illustrative example is constructed to demonstrate the calculation process and the computational effort required to perform the inverse algorithm.

In Chapter 3 we present our detailed formulation of the POMDP and inverse POMDP that are specific to the problem of finding and analyzing optimal time-dependent screening policies for detecting breast cancer in women of ages 25 to 100. We first provide background on breast cancer and breast cancer screening policies. The literature review discusses some breast cancer screening policy models and previous research on inverse problems in the healthcare area. Then, we construct a POMDP model of a regular breast cancer screening policy and discuss the application of the inverse algorithm from Chapter 2 to this POMDP model.

In Chapter 4, we start with the discussion of the experimental design of the inverse problem for the breast cancer screening policy application. Also, we introduce a performance metric, the expected number of mammograms, that is used to measure the effort of executing a particular screening policy. Next we discuss the data sources we used in our model and the numerical results that are obtained by applying the inverse algorithm.

Chapter 5 includes two major components to complete this research: the validation of the inverse algorithm and the sensitivity analysis. The purpose of the validation is to ensure that the imputed rewards from the inverse algorithm do make the designated time-dependent policy the optimal policy; the goal of the sensitivity analysis is to examine the change in the imputed rewards given some small fluctuations in the input. In Section 5.1, we first introduce the validation method. Then, we provide the validation results for some chosen age groups. Following the validation section, Section 5.2 includes details of the sensitivity analysis. We provide a sampling method for comprehensively exploring the impact of the small changes in the input that is specifically designed for the one-step transition probability matrix of MDPs/POMDPs. The result and discussions of the sensitivity analysis are provided after the description of sampling method. Finally in Chapter 6, we summarize the main findings of this research and we discuss directions for future reseach.

Chapter 2

The Inverse Algorithm for Time-Dependent Policies

2.1 Introduction

Tarantola [48] describes inverse modeling in the following manner:

> ... use of the actual results of some measurements of the observable parameters to infer the actual values of the model parameters.

Ahuja and Orlin [1] expand on Tarantola's words to compare inverse modeling problems with regular forward modeling problems; and they clearly explain the difference between the two problem types in the context of optimizing system performance:

> A typical optimization problem is a forward problem because it identifies the values of observable parameters (optimal decision variables), given the values of the model parameters (cost coefficients, right-hand side vector, and the constraint matrix). An inverse optimization problem consists of inferring the values of the model

parameters (cost coefficients, right-hand side vector, and the constraint matrix), given the values of observable parameters (optimal decision variables).

Hence, the answer to a forward problem is to give a decision maker recommendations of what to do, i.e., the "best" actions to be taken using the resources that are available in the current environment, i.e., the constraints. On the contrary, the solution to an inverse problem is to specify the properties of the constraints that cause a particular action to be preferred.

In the artificial intelligence area, *inverse reinforcement learning (IRL) problems* represent a rich diversity of inverse problems of much current interest. Reinforcement learning, from a computer science perspective, is the problem faced by an agent that must learn behavior through trial-and-error interactions with a dynamic environment [26]. An agent refers to the decision maker, i.e., a machine in the artificial intelligence area. An agent acts with a certain goal in a finite time horizon, and the horizon is divided into several discrete time periods which are called decision epochs.

A Markov chain is used to describe the status of an agent in a fully observable environment, which is known as the agent's "state" in the chain as that stochastic process evolves over time. The environment changes as a result of the action that the agent takes. The agent has to decide which action to take in order to achieve his goal at each decision epoch. A reward function is used to describe the actions quantitatively so that an agent can determine which action is the best in the long run. A set of actions forms a policy, and the optimal policy returns the largest expected reward that an agent can obtain over the time horizon of the problem. Thus, a reinforcement learning problem is built on the foundation of a Markov decision process (MDP) problem.

In the so-called grid world example, an agent starts in the bottom-left (southwest) square in a five-by-five grid world with the goal to go to the top-right (northeast) square, where he can

obtain a reward of one unit. Each square is a state in a Markov chain, and the agent's action choices are to move to the square above, below, to the right of, or to the left of the square he is in currently. The agent needs to select an action in each square, i.e., to form a policy that specifies an action for each state, so that he can maximize the expected reward.

Sometimes, the agent has to learn the optimal policy from an expert who does not offer an explicit reward function that the agent can use to calculate the best action. An expert here means someone who knows about the environment and how to act in different situations (states). Normally, an expert would first demonstrate how to make the best decision in different situations so that the agent can learn how to react within the dynamic environment from the expert. For example, when we first learn how to drive, an adult, normally a parent, would take us for a ride and show us how to react to various traffic conditions. Our parent's reward function is not clearly defined, but we can learn how to react from the parent's demonstration, with the goal of driving to our destination. Thus, the parent is the expert in this case. Therefore, learning from the expert reinforcement learning problem is also called "apprenticeship learning."

Regardless of whether the agent has to learn from trial-and-error or from an expert, the reward function is not clearly and completely specified. Therefore, researchers in the reinforcement learning area face the following problem: how to use the observed policy to infer the actual reward function; and this is the IRL problem.

Inverse reinforcement learning was first defined by Stuart Russell [41] as follows:

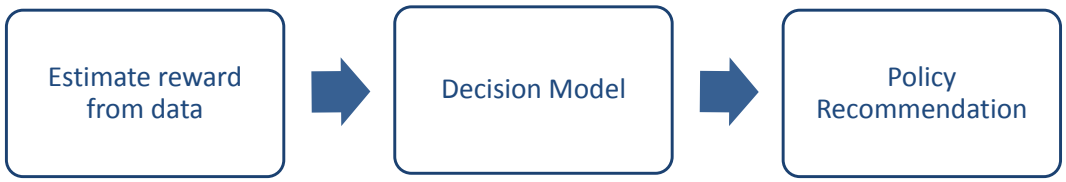
Given 1) measurements of an agent's behaviour over time, in a variety of circumstances; 2) measurements of the sensory inputs to that agent; and 3) a model of the physical environment (including the agent's body)

Determine the reward function that the agent is optimizing.

Therefore, the purpose of the IRL problem is to reconstruct the reward function from a policy as accurately as possible. In the grid world example, we see that an inverse reinforcement learning problem is to reconstruct the reward function given the action an agent would take in each square. In Chapter 3 we discuss an application of IRL to the analysis of breast cancer screening policies; and in this context, we seek to impute the "value" (for example, utility) of each of the possible health states for an individual woman from the policy that specifies the schedule by which she is screened for breast cancer.

The IRL definition above can be extended to partially observable Markov decision processes (POMDPs), which, as implied by the name, relax the perfect-state-information assumption of MDPs, i.e., the agent has a probabilistic information of the status of the process. An inverse POMDP is different from a forward POMDP. The rewards/costs of a forward POMDP are estimated from available data. Given a known reward for being in each state of a POMDP, we seek to identify the best action to take for each state that the process visits and at each decision epoch in the time horizon for the problem at hand. By contrast, an inverse POMDP seeks to estimate the rewards/costs that make a specified policy dominate all other policies. Accordingly, in an inverse POMDP, different policies should generate different sets of rewards/costs that make the specified policy better than all other policies.

Figure 2.1 illustrates the three-step work flow of a forward MDP/POMDP compared with that of an inverse MDP/POMDP. As the arrows show, the flow in Figure 2.1(a) is from left to right, while the flow in Figure 2.1(b) is from right to left. The first step in a forward problem, as shown in Figure 2.1(a), is to first estimate the rewards from available data sets. Then, we use the result from the first step as the input of the decision model, whose solution yields a policy, to recommend the decision maker. The inverse problem, as shown in Figure 2.1(b), works the opposite way. It assumes the decision maker follows a particular policy or observes a policy from a demonstration. The policy serves as the input of the inverse decision model. The goal of



(a) Forward Problem

(b) Inverse Problem

Figure 2.1: The forward and inverse problem work flows

the inverse decision model is to find the imputed rewards that make the followed policy better or not worse than all the other possible policies.

When solving any inverse problem, degeneracy is an important issue to consider. Degeneracy in the solution(s) of an inverse problem means that, in contrast to a forward problem, there may be multiple solutions to an inverse problem corresponding to a given input for the inverse problem—that is, the mapping from input to outputs for an inverse problem may not be one-to-one. For example, a given combination of rewards usually leads to only one policy recommendation, while a given policy recommendation can result from many different combinations of rewards.

POMDPs are widely used in healthcare decision making due to the indirectly observable nature of a patient's health status. However, a policy acquired from a forward POMDP may

be difficult to implement for the general public because the following difficulties may prevent accurate assessment of the belief of an individual's state of health: (a) diagnostic errors; (b) errors in recording the results of diagnostic procedures; and (c) unavailability of the individual's relevant medical history. Hence, many of the screening policy recommendations, such as the breast cancer screening policy of the American Cancer Society and the cervical cancer screening policy of the United States Preventive Services Task Force, depend more on timing than on a patient's health status. Therefore, the goal of this chapter is to develop an inverse algorithm for time-dependent policies, where a time-dependent policy refers to a rule for taking action that depends on the timing of the action before the patient's health status is symptomatic. Through this inverse algorithm, we can perform an analysis of the advantages and disadvantages of time-dependent policies given the necessary optimality conditions of time-dependent policies.

Mortality rate and quality-adjusted life-years (QALYs) are two commonly used reward metrics in healthcare decision making, especially in the disease modeling area. When using either of these rewards in a model, one important issue is that the reward may not be stationary. In other words, the reward changes with a patient's age. As a patient ages, these types of rewards normally change. Hence, our inverse algorithm also addresses the nonstationary reward issue.

When building a model to find the optimal screening policy, we must recognize that the sequence of events in a POMDP for disease screening might differ from the sequence of events in a typical POMDP. The purpose of a disease screening test is to gather information about the progression of disease, not to change the disease progression directly. Hence in a disease screening policy model, the screening result (observation) comes first, and then the transition in the patient's health state occurs according to the observation so that either the disease progresses naturally to the next time period given a normal screening result (i.e., the test result

suggests the patient does not have the disease), or the patient takes actions (normally receiving some kind of treatment) to change the disease's natural progression given an abnormal result (i.e., the test result suggests the patient may have the disease so that a more advanced test is necessary to reveal the patient's health status with regard to the disease), which will not be explicitly modeled in the disease screening model. Therefore, the formulation of such a model is slightly different as will be discussed in detail in Section 2.4.

The remainder of this chapter is organized as follows. Section 2.2 reviews the relevant literature on inverse algorithms for MDPs and POMDPs. In Section 2.3, we present a general discrete-time, finite-horizon POMDP with finite state and action spaces. In Section 2.4, we discuss the difference between a disease screening POMDP model and a general POMDP. Also, we present the formulation of a disease screening POMDP model. In Section 2.5, we present our algorithm for time-dependent policies. An example is presented to illustrate the inverse algorithm in Section 2.6. Finally in Section 2.7, we summarize our algorithm for imputing the reward function for a POMDP-based problem given a time-dependent policy for that problem; and we set the stage for applying this algorithm to breast cancer screening.

2.2 Literature Review

Inverse problems were first studied by geophysical scientists. Oftentimes, the model parameters are difficult, or even impossible, to measure accurately. For example, geophysical scientists have a model to calculate the gravitational field around a planet. The mass distribution of a planet is required to calculate the gravitational field. Geophysical scientists can estimate the gravitational field directly; but to do the inverse, inferring the distribution of mass from the estimated gravitational field, is not a one-to-one mapping problem. For the inverse problems, the geophysical scientists determine how to use the observable parameters, such as the

gravitational field, to estimate the model parameters, such as the mass distribution.

Geophysical scientists are also interested in modeling earthquake movements. To do so, the transmission time is a necessary but difficult-to-measure input. Geophysical scientists model the earthquake movement as a network problem, assuming that earthquakes travel along the shortest paths. A network can represent a large number of square cells by discretizing a geologic zone. The nodes in the network represent the cells, and the arcs connecting the nodes are used to indicate if two cells are adjacent. The costs associated with the arcs are the transmission times. Geophysical scientists can observe earthquakes and collect seismic perturbation data in order to infer the transmission times. Tarantola's books [47, 48] discuss the theories and applications of inverse problems in the geophysical sciences.

In the context of mathematical programming, researchers seek to solve an inverse optimization problem by perturbing the coefficients of the objective function in the corresponding forward problem as little as possible in order to make a given feasible solution become the optimal solution. Thus, researchers apply the L_p-norm concept to construct the inverse algorithm, where the L_p-norm, also known as the p-norm, of $x \in \mathbb{R}^n$ is defined by

$$\| x \|_p = \begin{cases} (| x_1 |^p + | x_2 |^p \ldots + | x_n |^p)^{\frac{1}{p}}, & \text{if} \quad 1 \leqslant p < \infty, \\ \max \{| x_i | : i = 1, \ldots, n\}, & \text{if } p = \infty. \end{cases} \tag{2.1}$$

Researchers construct the inverse problem, which is to minimize the difference between the original coefficients and the new coefficients under the L_p norm, where the L_1 norm, the L_2 norm and the L_∞ norm are most commonly used.

Burton and Toint [10, 11] first investigated the inverse shortest path problem. They proposed algorithms for searching the costs associated with the arcs in a weighted oriented graph given only a subset of the shortest paths in the graph. In their first article, the L_2 norm, also

known as the least-squares norm, is used to construct the inverse shortest path problem. Thus, the inverse shortest path problem is a quadratic programming problem. This inverse shortest path problem is to determine the costs by changing the coefficients in the shortest path problem as little as possible so that a given path between two designated vertices in the graph becomes the shortest path between those vertices. They assume the costs associated with the arcs are random variables that are nonnegative and independent of each other. The algorithm to solve the inverse problem is a specialization of the dual quadratic programming problem by Goldfarb and Idnani [19]. An example of such an inverse problem is to determine the travel cost, measured by the time delay, associated with each arc given the users' route choices in an unsaturated transportation network. The independent cost assumption was relaxed in the second article by Burton and Toint [11] to be suitable for more general shortest path problems. They modified their algorithm in order to solve more general inverse shortest path problems.

Zhang and Liu [53, 54] first explored inverse linear programming (LP) problems. Their goal is to minimize the L_1-norm of the difference between the new and old coefficients in the objective function so that a given feasible solution of the forward LP becomes the optimal solution. In the first article, they show the inverse problem under the L_1 norm of an LP problem is also an LP problem. They construct their algorithm based on the optimality conditions of an LP problem for two special problem classes: minimum cost flow problems and assignment problems. In the second article by Zhang and Liu [54], they extend their previous work to binary LP problems. Moreover, they consider inverse LP problems under the L_∞ norm for the same special problem classes.

Ahuja and Orlin [1] also studied inverse LP problems. They also construct the inverse problems with a goal that is similar to the goal of Zhang and Liu — namely, to minimize the L_p-norm of the difference between the new and old coefficients in the objective function so that a given feasible solution of the forward LP becomes the optimal solution. Extending

Zhang and Liu's work [53, 54], Ahuja and Orlin clarify and apply the relationship between the dual solution of the forward LP problem and the inverse problem for more general types of LP problems. They formulate the inverse LP problem under the L_1 norm as well as under the L_∞ norm. They show the following,

- The inverse problem under the L_1 norm (or the L_∞ norm) of an LP problem is also an LP problem;
- Solving the inverse problem under the L_1 norm with unit weights in a shortest path problem, an assignment problem, or a minimum-cut problem is the same as solving the corresponding forward problems; for the problem with nonunit weights, the inverse problem under the L_1 norm becomes a minimum-cost flow problem;
- The inverse problem under the L_1 norm of a minimum-cost flow problem (with unit cost) is similar to a (unit-capacity) minimum-cost flow problem;
- For a shortest path, assignment, or minimum-cost flow problem with unit cost, the inverse problem under the L_∞ norm becomes a minimum mean-cycle problem. With nonunit cost, the inverse problem under the L_∞ norm becomes a minimum cost-to-time ratio cycle problem; and
- For a linear polynomially solvable problem using the ellipsoid algorithm, the inverse problem under the L_1 (or L_∞) norm is also polynomially solvable.

The survey paper by Heuberger [21] examined several inverse combinatorial optimization problems and different algorithms to solve those problems.

In terms of machine learning, inverse reinforcement learning problems were initially studied by Ng and Russell [34]. They presented LP algorithms to seek the optimal rewards of MDPs

with state-dependent policies for finite and infinite state spaces. They utilize Bellman's principle of optimality and the Bellman equation to construct their inverse algorithm. For finite-state-space MDPs, they provide a characterization of the solution set that is sufficient to ensure a given policy is optimal — namely, that a column vector with nonnegative elements is obtained by computing the usual matrix product of the following: (i) the difference between the transition probability matrix for the optimal action and the transition probability matrix for any other action; (ii) the matrix inverse of the difference between the identity matrix and the transition probability matrix for the optimal action; and (iii) the reward function of this given policy. Then, a heuristic method (maximizing the distance between the optimal policy and all the other policies, i.e., maximizing the L_1 norm of the difference between the optimal policy and all the other policies) is used to address the degeneracy problem. For an infinite-state-space MDP, the reward function is assumed to be in the form of a linear combination of a collection of basis functions that are used to describe the state space, such as Gaussian-shaped basis functions approximating a position in a two-dimensional Euclidean space. Ng and Russell propose an heuristic method similar to their heuristic for inverse finite-state MDPs to address the degeneracy issue in inverse infinite-state MDPs. The proposed LP algorithm is used to maximize the difference between the value of the best policy and the value of the second-best policy for a large number of potential states that the process might enter.

Ramachandran and Amir [39] incorporate Bayesian inference concepts into Ng and Russell's inverse reinforcement learning algorithms to reconstruct the reward function and prove that Ng and Russell's inverse reinforcement learning algorithms are special cases of the Bayesian inverse reinforcement learning algorithm formulated in [39]. The main concept is to use the expert's behavior as the prior to derive a probability distribution over the space of reward functions. Similar to Ng and Russell's algorithms, the reward function is described by a distribution depending on the structure of the state space. Ramachandran and Amir assume that the expert's

goal is to maximize the overall reward, and the expert's actions are consistent and stationary whenever the expert visits a given state. The prior distribution on the reward function is chosen from a known family of prior distributions; and the likelihood distribution, the conditional probability distribution of an expert's action in a particular state given the reward function, follows an exponential distribution. Ramachandran and Amir construct the posterior probability distribution of the reward function from the prior and the likelihood function. The algorithm employs the Markov chain Monte Carlo algorithm for sampling from the prior distribution over the space of reward functions.

An algorithm for the partially observable environment is proposed by Choi and Kim [14]. To solve the continuous belief state space issue, they introduce the finite-state controller concepts from graph theory into Ng and Russell's inverse reinforcement learning algorithm. A policy can be converted into a graph in which some nodes are associated with actions and some directed arcs are associated with observations. Hence, the finite-state controller graph is used to sort the reachable belief states into nodes so that the continuous belief state space can be approximated by a finite number of nodes and arcs. To search for the reward function, Choi and Kim propose a linear program similar to Ng and Russell's algorithm. The algorithm seeks to maximize the difference between the total value of the nodes generated from the optimal policy and the total value of the nodes generated from the other policies. The policy optimality condition for the finite state controller forms the constraint set. A survey paper by Zhifei and Joo [57] provides reviews and comparisons of the original inverse learning algorithm of Ng and Russell [34] and close variants of the latter algorithm.

Our study in this chapter differs from the above referenced work in that, to the best of our knowledge, we have developed the first algorithm that is focused on recovering the reward function for POMDPs with nonstationary (i.e., time-dependent) policies and nonstationary immediate rewards. One may argue that time can be modeled as a part of the POMDP state space.

However, the computational expense of a state-dependent policy for MDPs with stationary transition probabilities is already significant. Adding time as a part of the state space definition is not very attractive when we consider the computational complexity of this approach. Also, the inverse reinforcement learning algorithms are designed more for building an agent who can perform a task by mimicking the expert's behavior. Our study concentrates more on recovering the reward function so that one can compare different policies.

2.3 The Finite-State POMDP

We begin with the definition of an MDP to introduce the idea of a POMDP with finite (discrete) state and action spaces, a finite time horizon H, and the discrete set $\mathbb{T} = \{t : t = 0, \ldots, H\}$ of decision epochs, where 0 and H denote the initial and final decision epochs in the MDP, respectively.

A finite-state MDP is defined by a 4-tuple $(\Omega, \mathcal{A}, \mathcal{T}, \mathcal{R})$ where: $\Omega = \{i : i = 1, \ldots, \nu\}$ is the finite state space of the core process $\{x^{\,t} : t \in \mathbb{T}\}$. Note that for ease of exposition, we assume that first ρ states in the state space are transient states where $\rho \leqslant \nu$. Let $\mathcal{A} = \{a_1, \ldots, a_\kappa\}$ denote the finite action space, which, for simplicity, is assumed to remain the same over the entire time horizon. The set

$$\mathcal{T} = \left\{\mathcal{T}_{i,j}^{t}(a) : i, j \in \Omega; a \in \mathcal{A}; 0 \leqslant t < H\right\} \tag{2.2}$$

is composed of all the relevant state transition functions, where

$$\mathcal{T}_{i,j}^{t}(a) = \Pr\left\{x^{\,t+1} = j \mid x^{\,t} = i \text{ and action } a \text{ is taken at decision epoch } t\right\} \tag{2.3}$$

denotes the probability of the core process moving from state i at decision epoch t into state j

at decision epoch $t+1$ when action a is taken at decision epoch t;

$$\mathcal{R} = \left\{ q_i^t(a) : i \in \Omega; a \in \mathcal{A}; 0 \leqslant t \leqslant H \right\} \tag{2.4}$$

is the reward function, where $q_i^t(a)$ denotes the immediate reward accrued if the process is in state i at decision epoch t and action a is taken.

Let x^t denote the state of the core process at decision epoch t. The stochastic process $\{x^t : t = 0, \ldots, H\}$ is assumed to satisfy the Markov property. In other words, the state in the next time period only depends on the current state and the action taken at the current decision epoch; the previous states and actions do not affect where the process will be in the future.

Consider an MDP with two states, $\Omega = \{i_1, i_2\}$, and two actions, $\mathcal{A} = \{a_1, a_2\}$, as an example. We assume that the transition function $\mathcal{T}_{i,j}^t(a)$, for $i, j \in \Omega$, and $a \in \mathcal{A}$, is stationary for simplicity, i.e., $\mathcal{T}^t$ does not change with time so the superscript t is suppressed in the following discussion. The following is a sample transition function written in a matrix form,

$$\mathcal{T}_{i,j}(a_1) = \begin{array}{c} \\ i_1 \\ i_2 \end{array} \begin{array}{c} \begin{array}{cc} i_1 & i_2 \end{array} \\ \begin{bmatrix} 0.3 & 0.7 \\ 0.8 & 0.2 \end{bmatrix} \end{array}, \qquad \mathcal{T}_{i,j}(a_2) = \begin{array}{c} \\ i_1 \\ i_2 \end{array} \begin{array}{c} \begin{array}{cc} i_1 & i_2 \end{array} \\ \begin{bmatrix} 0.5 & 0.5 \\ 0.3 & 0.7 \end{bmatrix} \end{array}. \tag{2.5}$$

If the process is currently in state i_1 and the action a_1 is taken, then at the end of the next time step, the process will enter state i_1 with probability 0.3 and state i_2 with probability 0.7. The summation of the transition probabilities from the same state should be equal to one, i.e., each row in a transition matrix should sum to one. For instance, $\mathcal{T}_{i_1,i_1}(a_1) + \mathcal{T}_{i_1,i_2}(a_1) = 0.3 + 0.7 = 1$.

A *MDP policy* is normally defined as a mapping φ from the state space Ω to the action

space $\mathcal{A}$,

$$\varphi : i \in \Omega \mapsto a = \varphi(i) \in \mathcal{A}. \tag{2.6}$$

In a forward MDP, the objective of a decision maker is to search for an optimal policy, $\widetilde{\varphi}$. Through this mapping, $\widetilde{\varphi}$, a decision maker can obtain the largest expected value of the reward over the time horizon.

Sometimes, an immediate reward associated with a specific action is bigger than the immediate rewards associated with the other actions. In other words, a specific action is more attractive for a decision maker to take at the current decision epoch t than all other actions. However, a decision maker should also consider all possible future rewards received from those states that the process might enter later. Hence, a decision maker might sacrifice the bigger reward now in exchange for a higher overall value accumulated over the entire time horizon of the problem at hand.

For example, suppose the immediate reward $q_i^t(a_1)$ of taking action a_1 in state i is bigger than another immediate reward $q_i^t(a_2)$ of taking action a_2 in state i at the current decision epoch t. However, taking action a_2 might cause the process to enter states with much higher rewards later, while taking action a_1 might cause the process to enter states with lower rewards in the future. Hence, the decision maker should take action a_2, instead of action a_1, to obtain a larger expected overall value.

For $i \in \Omega$ and $t \in \{0,\ldots,H\}$, the value function $V^t(i)$ of an MDP is defined to be the expected value of the rewards accumulated at decision epochs $t,\ldots,H$ by following a policy, $\varphi(\cdot)$, starting at decision epoch t when the core process is currently in the state i. To search for the optimal policy, $\widetilde{\varphi}(\cdot)$, is equivalent to finding the largest expected value obtained from the value function, $\widetilde{V}^t(i)$. Hence, the value function, $\widetilde{V}^t(i)$, to obtain the largest expected reward

starting from state i at decision epoch t can be defined via the Bellman equation:

$$\widetilde{V}^{t}(i) = \begin{cases} \max\limits_{a\in\mathcal{A}} \left\{ q_i^{\,t}(a) + \sum\limits_{j\in\Omega} \mathcal{T}_{i,j}^{t}(a)\, \widetilde{V}^{\,t+1}(j) \right\}, & \text{for } 0 \leqslant t < H, \\ q_i^{\,H}(a), & \text{for } t = H, \end{cases} \tag{2.7}$$

where $i \in \Omega$;

and if we take

$$\widetilde{V}^{H+1}(i) \equiv 0 \text{ for } i \in \Omega, \tag{2.8}$$

then the optimal policy $\widetilde{\varphi}$ is obtained from Eq. (2.7) as follows:

$$\widetilde{\varphi}(i) = \arg\max_{a\in\mathcal{A}} \left\{ q_i^{t}(a) + \sum_{j\in\Omega} \mathcal{T}_{i,j}^{t}(a)\, \widetilde{V}^{t+1}(j) \right\} \tag{2.9}$$

Eq. (2.7) shows that a discrete-time and finite-state MDP can be defined by the state transition function and the reward function.

In an MDP, the underlying Markov chain $\{x^{\,t} : t = 0, \ldots, H\}$ (the core process) is completely observable. On the other hand, in a POMDP, the core process is not directly observable so that its current state $x^{\,t}$ at decision epoch t cannot be known with certainty. Thus, a decision maker can only learn about the value of $x^{\,t}$ through a message, known as an *observation*, that the core process sends at each decision epoch, t. A POMDP is defined by a 6-tuple $(\Omega, \mathcal{A}, \mathcal{Z}, \mathcal{T}, \mathcal{O}, \mathcal{R})$ where:

- $\Omega, \mathcal{A}, \mathcal{T}, \mathcal{R}$ are defined similarly as for MDP;
- $\mathcal{Z} = \{\ell : \ell = 1, \ldots, \zeta\}$ is the finite observation space, and the time series of observations, also known as the observation process, $\{z^{\,t} : t = 1, \ldots, H\}$ starting at decision epoch 1 is a realization of a stochastic process related to the unobservable core process $\{x^{\,t} : t =$

$1,\ldots,H\}$ starting at decision epoch 1 such that there is a stochastic linkage between $x^{\,t}$ and $z^{\,t}$ which we will exploit for $t = 1,\ldots,H$, where $\zeta = |\,\mathcal{Z}\,|$ is the number of elements in $\mathcal{Z}$, and;

- The set $\mathcal{O} = \left\{ \mathcal{O}_{j,\ell}^{\,t} : j \in \Omega; \ell \in \mathcal{Z}; 0 \leqslant t \leqslant H-1 \right\}$ is composed of all the relevant observation probability mass functions, where

$$\mathcal{O}_{j,\ell}^{\,t} = \Pr\left\{ z^{\,t+1} = \ell \middle| x^{\,t+1} = j \right\} \tag{2.10}$$

denotes the conditional probability that at decision epoch $t+1$ the observation $z^{\,t+1} = \ell$ is made, given that at decision epoch $t+1$, the core process is in state $x^{\,t+1} = j$. Note that the observation function $\mathcal{O}_{j,\ell}^{\,t}$ only depends on the state $x^{\,t+1} = j$ the core process enters at decision epoch $t+1$ and the observation $z^{\,t+1} = \ell$ that is made at decision epoch $t+1$, i.e., the observation is independent of the action.

We extend the previous MDP example to a POMDP example, where the state space Ω and the action space $\mathcal{A}$ each have two elements so that we have two states $\Omega = \{i_1, i_2\}$ and two actions $\mathcal{A} = \{a_1, a_2\}$. For a POMDP, the observation, $z^{\,t} \in \mathcal{Z}$, is the one additional factor to consider compared with an MDP. Assuming two possible observations $\mathcal{Z} = \{\ell_1, \ell_2\}$, we see that $\mathcal{O}_{i_1,\ell_2}^{\,t}$ is the conditional probability that at decision epoch $t+1$ we will make the observation $z^{\,t+1} = \ell_2$ given that the core process is in the state $x^{\,t+1} = i_1$ at decision epoch $t+1$. Note carefully that

$$\mathcal{O}_{i_1,\ell_2}^{t} = \Pr\left\{ z^{t+1} = \ell_2 \middle| x^{t+1} = i_1 \right\} \tag{2.11}$$

does not depend on the previous state x^t of the core process or on the action a^t taken at decision epoch t that also influenced the way in which the core process made a transition into state i_1 at decision epoch $t+1$. The observation function can be also expressed in a matrix form, such as

the following,

$$\mathcal{O}_{j,\ell} = \begin{array}{c} \\ i_1 \\ i_2 \end{array} \begin{array}{c} \begin{array}{cc} \ell_1 & \ell_2 \end{array} \\ \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix} \end{array}. \tag{2.12}$$

We assume that, in this example, the observation probability mass functions are stationary for simplicity; and thus we suppress the dependence on t in the following discussion of this example.

Since the core process $\{x^t : t = 0, \ldots, H\}$ is not directly observable, we have only some "belief" about the state of the core process, i.e., a probability distribution over the state space Ω. Let ${\pi_i}^t \in [0,1]$ denote $\Pr\{x^t = i\}$, the probability (belief) that the core process is in state i at decision epoch t. Therefore, we can use a vector $\boldsymbol{\pi}^t = \left[{\pi_1}^t, {\pi_2}^t, {\pi_3}^t, \ldots {\pi_v}^t\right]$ to represent the complete set of such beliefs at decision epoch t. The vector, $\boldsymbol{\pi}^t$, is a sufficient statistic that summarizes all of the information necessary for decision making at decision epoch t [7, 45]. Hence, $\boldsymbol{\pi}^t$ is also called the *belief state* or the *information vector* at decision epoch t. For example, $\boldsymbol{\pi}^t = [0.2, 0.8]$ means that the process is believed to be in states i_1 and i_2 with probabilities 0.2 and 0.8 respectively, while $\boldsymbol{\pi}^t = [1, 0]$ means that the process is believed to be in state i_1 with probability one.

The *Bayesian updating formula* is used to represent our belief about the transition of the core process being in state i at decision epoch t and moving into state j at decision epoch $t+1$ given that action a is taken at decision epoch t and we make the observation, $z^{t+1} = \ell$ at time

$t+1$,

$$\begin{aligned}\pi_j^{t+1}(a,\ell) &= \Pr\left\{ x^{\,t+1} = j \middle| \begin{array}{l}\text{action } a \text{ is taken at decision epoch } t \text{ and} \\ \\ \text{the observation at time } t+1 \text{ is } z^{\,t+1} = \ell\end{array} \right\} \\ &= \frac{\sum_{i\in\Omega} \pi_i{}^t \mathcal{T}_{i,j}^t(a)\, \mathcal{O}_{j,\ell}^{\,t}}{\sum_{i\in\Omega}\sum_{k\in\Omega} \pi_i{}^t \mathcal{T}_{i,k}^t(a)\, \mathcal{O}_{k,\ell}^{\,t}} \qquad (2.13) \\ &\qquad \text{for } j\in\Omega, a\in\mathcal{A}, \ell\in\mathcal{Z}, \text{ and } t = 0,1,\ldots,H-1.\end{aligned}$$

Assume the belief state $\pi^{\,t}$ is $[0.2, 0.8]$ at decision epoch t and the observation after taking action a_1 is ℓ_1. The belief that the state is i_2 at decision epoch $t+1$ given action a_1 is taken at decision epoch t and the observation ℓ_1 is made at decision epoch $t+1$ is calculated from Eq. (2.5), Eq. (2.12) and Eq. (2.13) as the following,

$$\begin{aligned}\pi_{i_2}^{t+1}(a_1,\ell_1) &= \frac{\pi_{i_1}{}^t \mathcal{T}_{i_1,i_2}^t \mathcal{O}_{i_2,\ell_1} + \pi_{i_2}{}^t \mathcal{T}_{i_2,i_2}^t \mathcal{O}_{i_2,\ell_1}}{\pi_{i_1}{}^t \mathcal{T}_{i_1,i_1}^t \mathcal{O}_{i_1,\ell_1} + \pi_{i_1}{}^t \mathcal{T}_{i_1,i_2}^t \mathcal{O}_{i_2,\ell_1} + \pi_{i_2}{}^t \mathcal{T}_{i_2,i_1}^t \mathcal{O}_{i_1,\ell_1} + \pi_{i_2}{}^t \mathcal{T}_{i_2,i_2}^t \mathcal{O}_{i_2,\ell_1}} \\ &= \frac{0.2\times 0.7\times 0.5 + 0.8\times 0.2\times 0.5}{0.2\times 0.3\times 0.9 + 0.2\times 0.7\times 0.5 + 0.8\times 0.8\times 0.9 + 0.8\times 0.2\times 0.5} \\ &\approx 0.1185.\end{aligned}$$

A well-known theory [3, 5, 32, 40, 45, 42] ensures that a POMDP can be converted into an equivalent completely observable MDP with a continuous state space Π. Let $\mathbb{R}$ denote the set of all real numbers, and let $\mid \Omega \mid = \nu$ denote the number of states in the state space of the original POMDP so that $\mathbb{R}^\nu$ denotes ν-dimensional Euclidean space. The equivalent completely observable MDP has the state space,

$$\Pi \equiv \left\{ \pi \in \mathbb{R}^\nu \middle| \sum_{i\in\Omega} \pi_i = 1, 1 \geq \pi_i \geq 0, i \in \Omega \right\}. \qquad (2.14)$$

Along with the previous POMDP example with two states and two actions, the equivalent MDP state space is all the possible combinations of the probabilities, π_{i_1} and π_{i_2} of being in states i_1 and i_2, respectively, with the normalization requirement, $\pi_{i_1} + \pi_{i_2} = 1$.

A POMDP policy is defined as a mapping from the state space and the discrete time set to the action space,

$$\varphi : (\pi^{\,t}, t) \in \Pi \times \mathbb{T} \mapsto a = \varphi(\pi^{\,t}, t) \in \mathcal{A}. \tag{2.15}$$

When the time reaches the last decision epoch H, the *terminal reward*, ${q_i}^H$, associated with being in the core state i, for every $i \in \Omega$, is received, because (a) there is no decision to make at the last decision epoch, and (b) there are no future rewards beyond time H that must be taken into account. Hence, the value function at time H is merely the expected value of the immediate reward taken over the probability distribution $\pi^{\,H}$ on the state space, Ω:

$$V^{\,H}(\pi^{\,H}) = \sum_{i \in \Omega} {\pi_i}^H {q_i}^H \text{ for every } \pi^{\,H} \in \Pi. \tag{2.16}$$

For decision epoch $t = H-1, H-2, \ldots, 0$, the value function associated with the equivalent MDP being in belief state $\pi^{\,t}$ and the action a being taken is defined recursively as follows:

$$\begin{aligned} V^{\,t}(\pi^{\,t}, a) = & \sum_{i \in \Omega} {\pi_i}^t {q_i}^t(a) \\ & + \sum_{i \in \Omega} \sum_{j \in \Omega} \sum_{\ell \in \mathcal{Z}} {\pi_i}^t \mathcal{T}_{i,j}^t(a)\, \mathcal{O}_{j,\ell}^t {V_j}^{t+1} \left({\pi_j}^{t+1}(a, \ell), \varphi(\pi^{\,t+1}, t+1) \right), \\ & \text{for every } \pi^{\,t} \in \Pi, a \in \mathcal{A}, \text{ and } t \in \mathbb{T} \setminus \{H\}, \end{aligned} \tag{2.17}$$

where the auxiliary functions $\{V_i^t(\pi_i^t, a^t) : i \in \Omega, t \in \mathbb{T}, a^t \in \mathcal{A}\}$ are defined recursively as fol-

lows:

$$V_i^{\,H}(\boldsymbol{\pi}^{\,H}, a^{\,H}) \equiv \pi_i^{\,H} q_i^{\,H} \text{ for } i \in \Omega \text{ and } a^{\,H} \in \mathcal{A}, \tag{2.18a}$$

$$\begin{aligned} V_i^t(\boldsymbol{\pi}_i^t, a^t) \equiv\ & \pi_i^{\,t} q_i^{\,t}(a^t) + \\ & \pi_i^{\,t} \sum_{j\in\Omega} \sum_{\ell\in\mathcal{Z}} \mathcal{T}_{i,j}^{\,t}(a^t) \mathcal{O}_{j,\ell} V_j^{t+1}\left(\boldsymbol{\pi}_j^{t+1}\left(a^t,\ell\right), \boldsymbol{\varphi}\left(\boldsymbol{\pi}^{t+1}(a^t,\ell), t+1\right)\right) \\ & \text{for } \boldsymbol{\pi} \in \Pi, i \in \Omega, a^t \in \mathcal{A}, \text{ and } t \in \mathbb{T} \setminus \{H\}, \end{aligned} \tag{2.18b}$$

where the ν-dimensional state vector $\boldsymbol{\pi}^{\,t+1}(a^t,\ell)$ in Eq. (2.18b) has the form

$$\boldsymbol{\pi}^{\,t+1}(a^t,\ell) \equiv \left[\pi_1^{\,t+1}(a^t,\ell), \ldots, \pi_\nu^{\,t+1}(a^t,\ell)\right] \text{ for } a^t \in \mathcal{A}, \text{ and } \ell \in \mathcal{Z}, \tag{2.19}$$

and for each $j \in \{1,\ldots,\nu\}$, the jth component of $\boldsymbol{\pi}^{\,t+1}(a^t,\ell)$ is the conditional probability $\pi_j^{\,t+1}(a^t,\ell)$ in the related POMDP as given by the Bayesian updating formula Eq. (2.13). Note that the value function at decision epoch t, $V^{\,t}(\boldsymbol{\pi}^{\,t}, a^t)$, is a weighted average over the core Markov chain state space Ω of the immediate rewards $\{q_i^t(a^t), i \in \Omega\}$ at the current decision epoch t and the auxiliary functions $\left\{V_j^{t+1}\left(\boldsymbol{\pi}_j^{t+1}\left(a^t,\ell\right), \boldsymbol{\varphi}\left(\boldsymbol{\pi}^{t+1}(a^t,\ell), t+1\right)\right) : j \in \Omega, \ell \in \mathcal{Z}\right\}$ at the future decision epoch $t+1$. Each core state i in the POMDP for $i \in \Omega$ carries the weight π_i^t for the immediate reward; and the weights for the future expected value involve the products of the following three components for each $j \in \Omega$ and $\ell \in \mathcal{Z}$:

- π_i^t, the probability of being in the core state $x^t = i$ at decision epoch t;
- $\mathcal{T}_{i,j}^{\,t}(a^t)$, the probability that the process makes a transition to state $x^{t+1} = j$ at decision epoch $t+1$ given $x^t = i$ and the action a^t at decision epoch t; and
- $\mathcal{O}_{j,\ell}^{\,t}$, the probability that the observation $z^{t+1} = \ell \in \mathcal{Z}$ is received given that the process enters state $x^{t+1} = j$ at decision epoch $t+1$.

For a forward POMDP, the objective of the decision maker is the same as it is for a forward MDP: to search for an "optimal" policy. However, the decision maker needs to also consider the observation $z^{\,t} \in \mathcal{Z}$ at decision epoch t for $t = 1, \ldots, H$, in a forward POMDP. To formulate the optimal value function $\widetilde{V}^t(\pi^t)$ at decision epoch t, where $t \in \{1, \ldots, H\}$, for the belief state $\pi^{\,t} \in \Pi$, we define the auxiliary functions $\{\widetilde{V}_i^t(\pi_i^t) : i \in \Omega, t = 0, 1, \ldots, H, H+1\}$ recursively as follows:

$$\widetilde{V}_i^{H+1}(\pi_i) \equiv 0 \text{ for } \pi \in \Pi, \tag{2.20}$$

and

$$\widetilde{V}_i^H(\pi_i^H) \equiv \pi_i^H q_i^H \text{ for } i \in \Omega, \tag{2.21a}$$

$$\begin{aligned} \widetilde{V}_i^t(\pi_i^t) \equiv{}& \pi_i^t q_i^t\left(\widetilde{\varphi}(\pi^t, t)\right) + \\ & \pi_i^t \sum_{j\in\Omega}\sum_{\ell\in\mathcal{Z}} \mathcal{T}_{i,j}^t\left(\widetilde{\varphi}(\pi^t, t)\right) \mathcal{O}_{j,\ell}^t \widetilde{V}_j^{t+1}\left(\pi_j^{t+1}\left(\widetilde{\varphi}(\pi^t, t), \ell\right)\right) \\ & \text{for } i \in \Omega, \text{ and } t \in \mathbb{T} \setminus \{H\}, \end{aligned} \tag{2.21b}$$

where for $t = H-1, H-2, \ldots, 1, 0$, the optimal policy

$$\widetilde{\varphi} : (\pi^{\,t}, t) \in \Pi \times \mathbb{T} \mapsto a = \widetilde{\varphi}(\pi^{\,t}, t) \in \mathcal{A}, \tag{2.22}$$

is defined recursively along with the auxiliary functions $\{\widetilde{V}_i^t(\pi_i^t)\}$ as follows:

$$\begin{aligned} \widetilde{\varphi}\left(\pi^{\,t}, t\right) = \arg\max_{a\in\mathcal{A}} \Bigg\{ & \sum_{i\in\Omega} \pi_i{}^t q_i{}^t(a) \\ & + \sum_{i\in\Omega}\sum_{j\in\Omega}\sum_{\ell\in\mathcal{Z}} \pi_i{}^t \mathcal{T}_{ij}^t(a)\, \mathcal{O}_{j,\ell}^{\,t} \widetilde{V}_j{}^{t+1}\left(\pi_j{}^{t+1}(a, \ell)\right) \Bigg\}. \end{aligned} \tag{2.23}$$

Observe that the auxiliary functions $\left\{\widetilde{V}_i^H(\pi_i^H) : i \in \Omega\right\}$ at decision epoch H are used to define

the optimal policy $\widetilde{\varphi}(\pi^{H-1}, H-1)$ at decision epoch $H-1$ by applying Eq. (2.23) with $t = H-1$; and then we apply Eq. (2.21b) with $t = H-1$ to compute the auxiliary function $\widetilde{V}_i^{H-1}(\pi_i^{H-1})$. Clearly, this process of alternately applying Eq. (2.23) and Eq. (2.21b) in a joint recursion enables us to compute both the complete optimal policy Eq. (2.22) and the complete set of auxiliary functions Eq. (2.21).

Note that, an optimal policy of a POMDP, $\widetilde{\varphi}\,(\pi^{\,t}, t)$, provides the decision maker a recommended action given the belief state $\pi^{\,t}$ at a particular decision epoch t. In other words, even if a belief state $\pi^{\,t}$ is equivalent to a belief state $\pi^{\,t'}$, e.g., $\pi^{\,t} = \pi^{\,t'} = [0.2, 0.8]$, the recommended action might be action a at decision epoch t and a different action a' at another decision epoch t'.

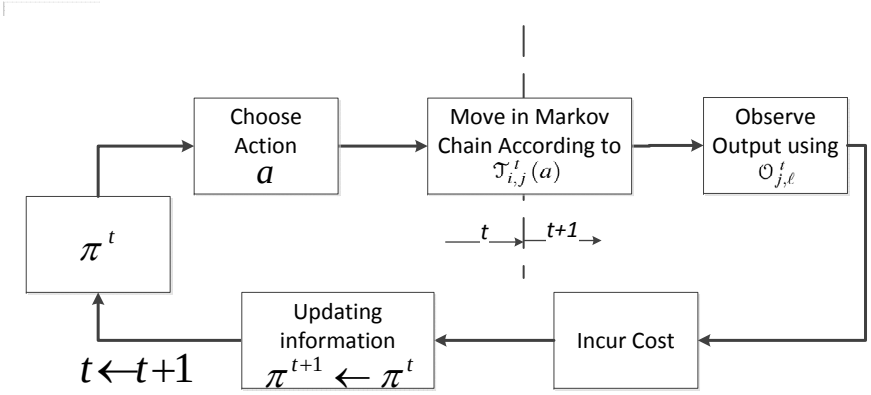
With this setup, the Bellman equation for the optimal value function $\widetilde{V}^t\,(\pi^t)$, where $t \in \{1, \ldots, H\}$ and $\pi^t \in \Pi$, has the form

$$\widetilde{V}^{\,t}(\pi^{\,t}) = \max_{a \in \mathcal{A}} \left\{ \sum_{i \in \Omega} {\pi_i}^t q_i^{\,t}\,(a) + \sum_{i \in \Omega} \sum_{j \in \Omega} \sum_{\ell \in \mathcal{Z}} {\pi_i}^t \mathcal{T}_{i,j}^{\,t}\,(a)\, \mathcal{O}_{j,\ell}^{\,t}\, \widetilde{V}_j^{\,t+1} \left(\pi_j^{\,t+1}\,(a, \ell) \right) \right\}. \tag{2.24}$$

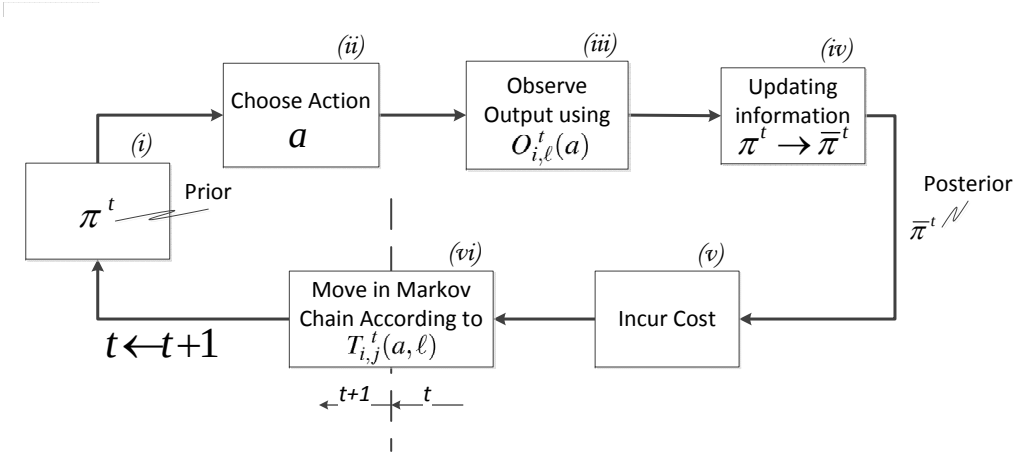
2.4 The POMDP for Disease Screening

As we mentioned in Section 2.1, the disease screening POMDP is different from a typical POMDP normally introduced in the literature. Figure 2.2 shows the two schematic flows of the decision processes. In a typical POMDP, shown as in Figure 2.2a [32], after the action a is chosen at decision epoch t, the process makes the transition to the next state x^{t+1} according to $\mathcal{T}_{i,j}^{\,t}(a)$, the conditional probability, Eq. (2.3), that $x^{t+1} = j$ given that $x^t = i$ and action a is taken at decision epoch t; then the decision maker receives the observation z^{t+1} according to $\mathcal{O}_{j,\ell}^{\,t}$, the conditional probability, Eq. (2.10) that $z^{t+1} = \ell$ given that $x^{t+1} = j$, in order to determine the

relevant incurred cost at decision epoch $t+1$. Finally in a typical POMDP, the associated belief state π^{t+1} at decision epoch $t+1$ is computed using the Bayesian updating formula Eq. (2.13) as depicted in the event labelled "Updating information $\pi^{t+1} \leftarrow \pi^{t}$" after the event labelled "Incur cost" in Figure 2.2a. Note that throughout our discussion of POMDPs, the terms "cost" and "reward" are used interchangeably.



(a) a POMDP flow from Monahan et al. [32]



(b) a POMDP flow for Disease Screening

Figure 2.2: The conventional POMDP [32] and the POMDP for Disease Screening

The "action" in a disease screening model is to "learn about the underlying core state." Hence, shown as in Figure 2.2b, after the action a is chosen (see the box labelled (*ii*)), the decision maker receives the observation z^t according to $O^t_{i,\ell}(a)$, the conditional probability that $z^t = \ell$ given that $x^t = i$ and action a is taken at decision epoch t (see the box labelled (*iii*)). Then as shown in the box labelled (*iv*), the belief state at decision epoch t is updated from the prior distribution $\boldsymbol{\pi}^t$ to the posterior distribution $\overline{\boldsymbol{\pi}}^{\,t}(a,\ell)$ that takes account of the action $a \in \mathcal{A}$ and the observation $z^t = \ell \in \mathcal{Z}$ at decision epoch t through a revised Bayesian updating formula (see Eq. (2.28a) below) based on the conditional probabilities $\left\{O^t_{i,\ell}(a) : i \in \Omega, a \in \mathcal{A}, \ell \in \mathcal{Z}\right\}$. The next step in the disease screening POMDP depicted in Figure 2.2b is to determine the relevant cost incurred at decision epoch t (see the box labelled (*v*)).

To advance to the next decision epoch $t+1$ in the disease screening POMDP, in box (*vi*) of Figure 2.2b, we use Eq. (2.30) below to compute $\pi^{t+1}_j(a,\ell)$, the conditional probability that $x^{t+1} = j$ given that at decision epoch t action $a \in \mathcal{A}$ is taken and observation $z^t = \ell \in \mathcal{Z}$ is made, in terms of the posterior probabilities $\{\overline{\pi}^t_i(a,\ell) : i \in \Omega\}$ (i.e., the components of the posterior distribution $\overline{\boldsymbol{\pi}}^t(a,\ell)$) and the one-step transition probabilities $\{T^t_{i,j}(a,\ell) : i \in \Omega\}$, where $T^t_{i,j}(a,\ell)$ is the conditional probability that $x^{t+1} = j$ given $x^t = i$ and at decision epoch t action a is taken and observation $z^t = \ell$ is made. Note that to compensate for the switch of the order of taking an observation and making a state transition in the core process, the transition probability matrix has to depend not only on the chosen action but also on the observation result, i.e., the transition probability is a conditional probability given the current state of the process, the chosen action, and the observation result. For example, a patient normally seeks treatment to change the disease's natural progression after receiving an abnormal test result while a patient will not take any action to change the natural progression given a normal test result. Also, the observation probability mass functions $\{O^t_{i,\ell}(a) : i \in \Omega, \ell \in \mathcal{Z}, a \in \mathcal{A}\}$ depend on not only the state of the process but also the taken action. Moreover, the updating formula Eq. (2.28a) in

our model differs accordingly from the Bayesian updating formula in Eq. (2.13).

Once the core process enters one of the absorbing states, the process will stay in the absorbing state for the remaining decision epochs. On the other hand, to maintain the belief state space, Eq. (2.14), the probabilities of being in one of the other transient states should take account of the fact that the core process enters one of the absorbing states. Hence, in addition to the result from the screening test, a special new observation must be added to the observation space $\mathcal{Z}$ to denote the patient's entry into one of the absorbing states. To properly describe the process, each absorbing state requires an observation which is added to the observation space to capture the fact that the process will stay in the particular absorbing state for the remaining decision epochs. When the process is in one of the transient states, the probability of receiving the observation that is associated with an absorbing state is equal to zero.

Given the discussion above, a disease screening POMDP is defined by the 6-tuple $(\Omega, \mathcal{A}, \mathcal{Z}, T, O, \mathcal{R})$ where:

- $(\Omega, \mathcal{A}, \mathcal{Z})$ are defined similarly as for a conventional POMDP;
- The observation set is defined as $O = \left\{ O^t_{i,\ell}(a) : i \in \Omega; \ell \in \mathcal{Z}; a \in \mathcal{A}; 0 \leqslant t \leqslant H \right\}$. The set O is defined similarly to the observation set in a conventional POMDP, but, instead of receiving the observation at decision epoch $t+1$ based on the state $x^{\,t+1}$, we receive the observation at decision epoch t based on the current state $x^{\,t}$. Hence, the observation probability mass functions in the set O are defined as follows,

$$O^t_{i,\ell}(a) = \Pr\left\{ z^{\,t} = \ell \middle| x^{\,t} = i \text{ and action } a \text{ is taken at decision epoch } t \right\},$$
$$\text{for } t = 0, 1, \ldots, H. \quad (2.25)$$

- The reward function is defined as

$$\mathcal{R} = \left\{ q_{i,\ell}^{t}(a) : i \in \Omega; a \in \mathcal{A}; \ell \in \mathcal{Z}; 0 \leqslant t \leqslant H \right\}, \tag{2.26}$$

  where the reward may not be stationary and depends on the state $x^{\,t} = i$, the taken action a, and the observation ℓ.

- The transition probability set $T = \left\{ T_{i,j}^{\,t}(a,\ell) : i,j \in \Omega, a \in \mathcal{A}, 0 \leqslant t \leqslant H \right\}$ is composed of all the relevant one-step state transition probabilities, where

$$T_{i,j}^{\,t}(a,\ell) = \Pr\left\{ x^{\,t+1} = j \middle| x^{\,t} = i, z^{\,t} = \ell, \text{ and action } a \text{ is taken at decision epoch } t \right\},$$
$$\text{for } t = 0,1,\ldots,H. \tag{2.27}$$

To match the 6-tuple definition above, we also need to calculate the posterior belief state $\overline{\pi}^{\,t}$ after taking action a at decision epoch t and then making the observation $z^{\,t} = \ell$. The Bayesian updating formula for a disease screening POMDP has the form

$$\begin{aligned} \overline{\pi}_i^{\,t}(a,\ell) &= \Pr\left\{ x^{\,t} = i \middle| z^{\,t} = \ell \text{ and action } a \text{ is taken at decision epoch } t \right\} \\ &= \frac{\pi_i^{\,t} O_{i,\ell}^{t}(a)}{\sum\limits_{k \in \Omega} \pi_k^{\,t} O_{k,\ell}^{t}(a)}, \quad \text{for } i \in \Omega, a \in \mathcal{A}, \ell \in \mathcal{Z}, t \in \{0,1,\ldots,H\}. \end{aligned} \tag{2.28a}$$

Eq. (2.28) is used to update the probability that the core process is in state i at decision epoch t, i.e., to update the belief state (or the information vector).

The basis for the Bayesian updating formula, Eq. (2.28a), warrants some additional explanation. Note that we assume the screening test does not change the disease progression and the action of not taking the screening test does not change the disease progression as well. In

other words, the probability of taking an action and the probability of being in the state i are independent of each other,

$$\Pr\Big\{\{x^t = i\} \bigcap \{\text{action } a \text{ is taken at decision epoch } t\}\Big\} = \Pr\{x^t = i\} \Pr\{\text{action } a \text{ is taken at decision epoch } t\}, \quad \text{for every } i \in \Omega, a \in \mathcal{A}, \text{ and } t \in \{0,1,\ldots,H-1\}. \tag{2.28b}$$

Hence, using Eq. (2.28b), we can rewrite Eq. (2.28a) as follows,

$$\begin{aligned}
\overline{\pi}_i^t(a,\ell) &= \Pr\Big\{x^t = i \Big| z^t = \ell, \text{ and action } a \text{ is taken at decision epoch } t\Big\} \\
&= \frac{\left[\begin{array}{c}\Pr\Big\{z^t = \ell \Big| x^t = i \text{ and action } a \text{ is taken at decision epoch } t\Big\} \\ \times \Pr\{\{x^t = i\} \bigcap \{\text{action } a \text{ is taken at decision epoch } t\}\}\end{array}\right]}{\left[\begin{array}{c}\sum\limits_{j\in\Omega} \Pr\Big\{z^t = \ell \Big| x^t = j \text{ and action } a \text{ is taken at decision epoch } t\Big\} \\ \times \Pr\{\{x^t = j\} \bigcap \{\text{action } a \text{ is taken at decision epoch } t\}\}\end{array}\right]} \\
&= \frac{\left[\begin{array}{c}\Pr\Big\{z^t = \ell \Big| x^t = i \text{ and action } a \text{ is taken at decision epoch } t\Big\} \\ \times \Pr\{x^t = i\} \Pr\{\text{action } a \text{ is taken at decision epoch } t\}\end{array}\right]}{\left[\begin{array}{c}\sum\limits_{j\in\Omega} \Pr\Big\{z^t = \ell \Big| x^t = j \text{ and action } a \text{ is taken at decision epoch } t\Big\} \\ \times \Pr\{x^t = j\} \Pr\{\text{action } a \text{ is taken at decision epoch } t\}\end{array}\right]} \\
&= \frac{\Pr\Big\{z^t = \ell \Big| x^t = i \text{ and action } a \text{ is taken at decision epoch } t\Big\} \Pr\{x^t = i\}}{\sum\limits_{j\in\Omega} \Pr\Big\{z^t = \ell \Big| x^t = j \text{ and action } a \text{ is taken at decision epoch } t\Big\} \Pr\{x^t = j\}}
\end{aligned} \tag{2.29}$$

Therefore, Eq. (2.28b) and Eq. (2.29) provide a complete derivation for the modified Bayesian

updating formula, Eq. (2.28a), in the disease screening model.

Given the posterior belief distribution $\overline{\pi}^t$, the process will make a transition to a new belief state π^{t+1} according to the chosen action and the observation result. Hence, the transition is made according to the following formula,

$$\begin{aligned}\pi_j^{t+1}(a,\ell) &= \Pr\left\{ x^{t+1} = j \,\middle|\, \begin{array}{l}\text{action } a \text{ is taken and observation } z^t = \ell \\ \text{is made at decision epoch } t\end{array} \right\}, \\ &= \sum_{i\in\Omega} \overline{\pi}_i^t(a,\ell) T_{i,j}^t(a,\ell), \quad \text{for } j \in \Omega, a \in \mathcal{A}, \ell \in \mathcal{Z}, t \in \{0,\ldots,H-1\}.\end{aligned} \tag{2.30}$$

To compute the optimal value function $\widetilde{V}^t(\pi^t)$ for all $t \in \mathbb{T}$ and $\pi^t \in \Pi$, we proceed along the lines of Eq. (2.15) – Eq. (2.24) with suitable modifications for our disease-screening POMDP. To maintain the optimal value function at the end of the time horizon, H, as specified by Eq. (2.16), it is convenient to take the final conditions

$$q_{i,\ell}^H(a) \equiv q_i^H \text{ for all } i \in \Omega, \ell \in \mathcal{Z}, a \in \mathcal{A}. \tag{2.31a}$$

We seek to define auxiliary functions $\{\widetilde{V}_i^t(\pi_i^t) : i \in \Omega; t = H+1, H, H-1, \ldots, 1, 0; a^t \in \mathcal{A}\}$ recursively as follows:

$$\widetilde{V}_i^{H+1}(\pi_i^{H+1}) \equiv 0 \text{ for all } i \in \Omega, \pi \in \Pi. \tag{2.31b}$$

$$\widetilde{V}_i^H(\pi_i^H) \equiv \pi_i^H q_i^H \text{ for } i \in \Omega, \tag{2.31c}$$

and

$$\widetilde{V}_i^t(\pi_i^t) \equiv \pi_i^t \sum_{\ell\in\mathcal{Z}} O_{i,\ell}^t\left(\widetilde{\varphi}(\pi^t,t)\right) q_{i,\ell}^t\left(\widetilde{\varphi}(\pi^t,t)\right) + \pi_i^t \sum_{\ell\in\mathcal{Z}}\sum_{j\in\Omega}\left\{O_{i,\ell}^t\left(\widetilde{\varphi}(\pi^t,t)\right) T_{i,j}^t\left(\widetilde{\varphi}(\pi^t,t),\ell\right)\widetilde{V}_j^{t+1}\left(\pi_j^{t+1}\left(\widetilde{\varphi}(\pi^t,t),\ell\right)\right)\right\} \tag{2.31d}$$

$$\text{for } i\in\Omega \text{ and } t = H-1, H-2, \ldots, 1, 0,$$

where the optimal policy Eq. (2.22) is defined recursively along with the auxiliary functions Eq. (2.31c) and Eq. (2.31d) according to

$$\widetilde{\varphi}^{\,t}(\pi^{\,t},t) = \arg\max_{a\in\mathcal{A}}\left\{\sum_{i\in\Omega}\sum_{\ell\in\mathcal{Z}} \pi_i{}^t O_{i,\ell}^t(a) q_{i,\ell}^t(a) + \sum_{i\in\Omega}\sum_{\ell\in\mathcal{Z}}\sum_{j\in\Omega} \pi_i{}^t O_{i,\ell}^t(a)\, T_{i,j}^{\,t}(a,\ell)\,\widetilde{V}_j{}^{t+1}\left(\pi_j{}^{t+1}(a,\ell)\right)\right\} \tag{2.32}$$

$$\text{for } \pi^t\in\Pi \text{ and } t = H, H-1, \ldots, 1, 0.$$

Note that the auxiliary functions $\{\widetilde{V}_i^H(\pi_i^H) : i\in\Omega\}$ at decision epoch H are used to compute the optimal policy $\widetilde{\varphi}(\pi^{H-1}, H-1)$ at decision epoch $H-1$ by applying Eq. (2.32); and then we apply Eq. (2.31d) with $t = H-1$ to compute the auxiliary functions $\{\widetilde{V}_i^{H-1}(\pi_i^{H-1}) : i\in\Omega\}$. Clearly, this process of alternately applying Eq. (2.32) and Eq. (2.31d) in a joint recursion enables us to compute both the complete optimal policy Eq. (2.32) and the complete set of the auxiliary functions Eq. (2.31d).

With this setup, the Bellman equation for the optimal value function has the form

$$\widetilde{V}^{\,t}(\pi^{\,t}) = \max_{a\in\mathcal{A}} \Bigg\{ \sum_{i\in\Omega}\sum_{\ell\in\mathcal{Z}} {\pi_i}^t O_{i,\ell}^t(a) q_{i,\ell}^t(a) + \sum_{i\in\Omega}\sum_{\ell\in\mathcal{Z}}\sum_{j\in\Omega} {\pi_i}^t O_{i,\ell}^t(a)\, T_{i,j}^t(a,\ell)\, \widetilde{V}_j^{\,t+1}\left({\pi_j}^{t+1}(a,\ell)\right) \Bigg\}, \tag{2.33}$$
$$\text{for every } \pi^{\,t} \in \Pi, \text{ and } t = 0,1,\ldots,H.$$

Observe when we take $t = H$ in the right-hand side of Eq. (2.33), we obtain

$$\begin{aligned} &\sum_{i\in\Omega}\sum_{\ell\in\mathcal{Z}} \pi_i^H O_{i,\ell}^H(a) q_i^H + \sum_{i\in\Omega}\sum_{\ell\in\mathcal{Z}}\sum_{j\in\Omega} {\pi_i}^t O_{i,\ell}^t(a)\, T_{i,j}^t(a,\ell)\, 0 \\ &= \sum_{i\in\Omega} \pi_i^H q_i^H \Bigg\{ \sum_{\ell\in\mathcal{Z}} O_{i,\ell}^H(a) \Bigg\} + 0 \\ &= \sum_{i\in\Omega} \pi_i^H q_i^H = V^H(\pi^H), \end{aligned} \tag{2.34}$$

which is the optimal value function at time H as required by Eq. (2.16). Note that, from the formulas for the optimal value function, Eq. (2.33), and the optimal policy, Eq. (2.32), we can see that the optimal value function is a linear combination of the immediate rewards, which provides a nice structure to develop the inverse algorithm.

To write the optimal value function, Eq. (2.33), and the optimal policy, Eq. (2.32), in a matrix form, a little trick is necessary to apply to the observation functions $\left\{ O_{i,\ell}^t(a) \right\}$ so that the dimensions in the formulas of the value function and the optimal policy will match. The trick is to introduce a diagonal matrix [32], $\mathbb{O}^{\,t}(a,\ell)$, where

$$\mathbb{O}_{i,j}^t(a,\ell) = \begin{cases} O_{i,\ell}^t(a) & i = j, \\ 0, & i \neq j, \end{cases} \qquad \text{for } i,j \in \Omega, a \in \mathcal{A}, \ell \in \mathcal{Z}, t \in \{0,1,2,\ldots,H\}. \tag{2.35}$$

Let

$$q^t(a,\ell) \equiv \begin{bmatrix} q^t_{1,\ell}(a,\ell) \\ q^t_{2,\ell}(a,\ell) \\ \vdots \\ q^t_{\nu,\ell}(a,\ell) \end{bmatrix} \text{ for all } a \in \mathcal{A}, \ell \in \mathcal{Z}, \text{ and } t = 0,1,\ldots,H \tag{2.36}$$

denote the $\nu \times 1$ reward vector given the action a and the observation result ℓ. Next we define the $\nu \times 1$ column vector

$$\widetilde{V}^t_c(\pi^t) = \begin{bmatrix} \widetilde{V}^t_1(\pi^t_1) \\ \widetilde{V}^t_2(\pi^t_2) \\ \vdots \\ \widetilde{V}^t_\nu(\pi^t_\nu) \end{bmatrix} \text{ for } t = 0,1,\ldots,H \text{ and } \pi^t \in \Pi \tag{2.37}$$

so that we can express the optimal value function Eq. (2.33) and the optimal policy Eq. (2.32) in the following compact matrix form:

$$\widetilde{V}^t(\pi^t) = \max_{a\in\mathcal{A}} \left\{ \pi^t \sum_{\ell\in\mathcal{Z}} \mathbb{O}^{\,t}(a,\ell) \left[q^t(a,\ell) + T^{\,t}(a,\ell) \widetilde{V}^{t+1}_c \left(\pi^{t+1}(a,\ell) \right) \right] \right\}, \tag{2.38}$$

$$\widetilde{\varphi}^t(\pi^t,t) = \arg\max_{a\in\mathcal{A}} \left\{ \pi^t \sum_{\ell\in\mathcal{Z}} \mathbb{O}^{\,t}(a,\ell) \left[q^t(a,\ell) + T^{\,t}(a,\ell) \widetilde{V}^{t+1}_c \left(\pi^{t+1}(a,\ell) \right) \right] \right\}, \tag{2.39}$$

where $T^t(a,\ell)$ denotes the $\nu \times \nu$ one-step transition probability matrix where (i,j) element $T^t_{i,j}(a,\ell)$ is defined by Eq. (2.27) for $i,j \in \Omega$, $a \in \mathcal{A}$, $\ell \in \mathcal{Z}$, and $t \in \mathbb{T}$.

2.5 Inverse Algorithm for Time-Dependent Policies

In this section, we will first introduce the concept of the heuristic inverse algorithm for time-dependent policies. The concept of the algorithm is to form a mathematical program that satis-

fies the requirement of a policy being the optimal policy by utilizing the property that a value function is a linear combination of the immediate rewards. This inverse algorithm seeks to impute the rewards/costs, $q_{i,\ell}^{t}(a)$, of a given time-dependent policy that is assumed to be optimal. However, combined with the unknowns, the rewards $q_{i,\ell}^{t}(a)$, and the belief states π^0 involved in the value function, the program is a nonlinear optimization problem. Luckily, the equivalent MDP state space Π forms a constraint of the program which provides a method to approximate the nonlinear program using several linear programs.

Let Φ denote the set of all possible time-dependent policies so that Φ has κ^H elements. From the definition of the optimal policy, Eq. (2.32), and the optimal value function, Eq. (2.33), we see that the value function for the optimal policy should be greater than or at least equal to the value function for all other possible policies. Therefore, given a time-dependent policy $\widetilde{\varphi} \in \Phi$ that is assumed to be optimal, we seek to impute appropriate values for the resulting optimal value function such that, for every other policy $\widehat{\varphi} \in \Phi \setminus \{\widetilde{\varphi}\}$, the difference between the respective imputed value functions for $\widetilde{\varphi}$ and $\widehat{\varphi}$ is nonnegative at beginning of the time horizon:

$$\widetilde{V}^0\left(\pi^0, \widetilde{\varphi}\right) - \widehat{V}^0\left(\pi^0, \widehat{\varphi}\right) \geqslant 0 \qquad \text{for all } \pi^0 \in \Pi, \widehat{\varphi} \in \Phi \setminus \{\widetilde{\varphi}\}. \tag{2.40}$$

Because the fundamental objective of the POMDP for disease screening and its equivalent observable MDP is to optimize $V^0(\pi^0, \varphi)$, the value function at decision epoch $t = 0$ over all $\pi^0 \in \Pi$ and over all possible time-dependent policies $\varphi \in \Phi$, it is reasonable to use the optimal value function $\widetilde{V}^0(\pi^0, \varphi)$ for all $\pi^0 \in \Pi$, $\varphi \in \Phi$ as the basic metric for distinguishing a presumed optimal policy $\widetilde{\varphi}$ from all its competitors $\widehat{\varphi} \in \Phi \setminus \{\widetilde{\varphi}\}$. Hence, the constraint Eq. (2.40) only considers the value function at the beginning of the entire time horizon.

Moreover, we seek to maximize over all initial belief states $\pi^0 \in \Pi$, the objective function defined as the sum taken over all nonoptimal policies $\widehat{\varphi} \in \Phi$ of the value-function differences

Eq. (2.40),

$$\max\left\{\sum_{\widehat{\varphi}\in\Phi\setminus\{\widetilde{\varphi}\}}\left[\widetilde{V}^0\left(\pi^0,\widetilde{\varphi}\right)-\widehat{V}^0\left(\pi^0,\widehat{\varphi}\right)\right]\,\middle|\,\begin{array}{l}\text{for all } \pi^0\in\Pi, \text{ and for}\\ \text{all reward functions Eq. (2.26)}\end{array}\right\}. \tag{2.41}$$

Maximizing the objective function Eq. (2.41) is designed to yield an imputed optimal value function that most clearly distinguishes the optimal time-dependent policy $\widetilde{\varphi}$ from all other time-dependent policies [34].

Formally, the nonlinear programming problem to be solved has the form,

$$\max_{\substack{\pi^0\in\Pi\\ \mathcal{R}}}\ \sum_{\widehat{\varphi}\in\Phi\setminus\{\widetilde{\varphi}\}}\left\{\widetilde{V}^0\left(\pi^0,\widetilde{\varphi}\right)-\widehat{V}^0\left(\pi^0,\widehat{\varphi}\right)\right\}, \tag{2.42a}$$

$$s.t.\quad \widetilde{V}^0\left(\pi^0,\widetilde{\varphi}\right)-\widehat{V}^0\left(\pi^0,\widehat{\varphi}\right)\geqslant 0 \qquad \text{for all } \pi^0\in\Pi,\widehat{\varphi}\in\Phi\setminus\{\widetilde{\varphi}\}, \tag{2.42b}$$

$$\sum_{i\in\Omega}\pi_i^t=1, \text{ for all } t\in\mathbb{T} \text{ and } \pi^t\in\Pi \tag{2.42c}$$

$$0\leqslant\pi_i^t\leqslant 1, \quad \text{for all } i\in\Omega, t\in\mathbb{T} \text{ and } \pi^t\in\Pi. \tag{2.42d}$$

Eq. (2.42b) is a constraint set whose number of constraints is one less than the number of available policies, i.e., κ^H-1. Also, we observe that Eq. (2.42c) and Eq. (2.42d) ensure that the objective function Eq. (2.42a) is maximized over the entire state space Π of the equivalent MDP.

Degeneracy is typically the main issue when solving an inverse optimization problem. In other words, a large set of reward functions $\mathcal{R}$ defined by Eq. (2.26) can lead to the same optimal policy. Similar to the approach of Ng and Russell [34] for handling an inverse MDP problem with a large state space, the objective function Eq. (2.42a) is intended to avoid the degeneracy issue. To do so, we use the maximization of the differences of the value function

results so that the value function of the specified policy is as far away as possible from the value function of all other possible policies.

The inverse algorithm seeks to solve for the rewards/costs of a given time-dependent policy that is assumed to be optimal. However, the rewards/costs existing in the value function combined with the belief state π^0 makes Eq. (2.42) a nonlinear program, which is not a computationally attractive characteristic for solving large-scale real-life applications. Fortunately, the structure of the belief state space Π as specified by Eq. (2.42c) and Eq. (2.42d) allows us to convert the nonlinear program into a linear program by discretizing the continuous belief state into a finite fixed grid. Hence, Eq. (2.42) has the following form for a fixed initial belief state π_x:

$$\max_{\mathcal{R}} \sum_{\widehat{\varphi}\in\Phi\setminus\{\widetilde{\varphi}\}} \left\{ \widetilde{V}^0(\pi_x,\widetilde{\varphi}) - \widehat{V}^0(\pi_x,\widehat{\varphi}) \right\}, \tag{2.43a}$$

$$s.t. \quad \widetilde{V}^0(\pi_x,\widetilde{\varphi}) - \widehat{V}^0(\pi_x,\widehat{\varphi}) \geqslant 0, \text{ for } \widehat{\varphi} \in \Phi \setminus \widehat{\varphi}, \tag{2.43b}$$

where $\pi_x = \pi^0 \in \Pi$. Let the grid $\mathscr{G} \subset \Pi$ denote a finite subset of selected initial belief states that in some sense "cover" the belief state space Π. For each $\pi_x \in \mathscr{G}$, we solve the LP Eq. (2.43) to compute the associated rewards $\{q^t_{i,\ell}(a)\}$; and then we must appropriately combine these rewards over all $\pi_x \in \mathscr{G}$, depending on the nature of the application at hand. For example, one can simply take the average imputed rewards,

$$\overline{q}^t_{i,\ell}(a) = \frac{1}{\left|\mathscr{G}\right|} \sum_{\pi^0_i:\pi^0\in\mathscr{G}} q^t_{i,\ell}(a) \text{ for } i \in \Omega, \ell \in \mathcal{Z}, t \in \mathbb{T}, \tag{2.44}$$

where $\left|\mathscr{G}\right|$ denotes the number of initial belief states in the grid $\mathscr{G}$, computed over all different initial belief states assuming that all initial belief states carry the same weight in the entire state

space Π.

The inverse algorithm imputes the rewards that make a particular policy the optimal policy. However, from the perspective of disease screening, the answer from the program, Eq. (2.42) may not be insightful, i.e., the imputed rewards may be operationally infeasible, which is shown in the example in Section 2.6. To resolve the operational infeasibility problem, the "age group α" concept is introduced. The time horizon is subdivided into several age groups, each having the length $\mathcal{L}$. For example, in the breast cancer screening policy application in Chapter 3, the length $\mathcal{L}$ is equal to 10 decision epochs, i.e., five years with six months between the decision epochs. Within an age group, an equality constraint involving the unknowns at different decision epochs is introduced into the inverse program, Eq. (2.42), i.e.,

$$q_{i,\ell}^{\alpha(t)}(a) = q_{i,\ell}^{t}(a) = q_{i,\ell}^{t+1}(a) = \cdots = q_{i,\ell}^{t+\mathcal{L}-1}(a), \quad i \in \Omega, \ell \in \mathcal{Z}, a \in \mathcal{A}, \tag{2.45a}$$

$$\alpha = \alpha(t) = \left\lceil \frac{t}{\mathcal{L}} \right\rceil + 1, \qquad \alpha = 1, 2, \ldots, \frac{H}{\mathcal{L}}, t \in \{0, 1, \ldots, H-1\}. \tag{2.45b}$$

In Chapter 3, 15 age groups are considered given the length of each age group is ten decision epochs, $\mathcal{L} = 10$. Also, applying the age group concept is similar to the conventional method of the reward setup in a forward POMDP.

However, the equality constraint Eq. (2.45) is a very strong assumption, which leads to the possibility of the infeasibility as found in the numerical example shown in Section 2.6. Hence, a robust idea is used to avoid the infeasibility problem that the equality constraint brings. The robust method is to relax the equality constraint with some tolerance $\varepsilon \geqslant 0$. We propose to use the unknowns at the beginning of each age group as the base values and to allow some deviation of maximum magnitude ε from those base values at different decision epochs in the

same age group as follows:

$$q^{\tau}_{i,\ell}(a) - \varepsilon \leqslant q^{\beta}_{i,\ell}(a) \leqslant q^{\tau}_{i,\ell}(a) + \varepsilon, \text{ for } \beta \in \{\tau+1, \tau+2, \ldots, \tau+\mathcal{L}-1\}, \tag{2.46a}$$

$$\tau = \mathcal{L}\,(\alpha(t)-1)+1, \quad t \in \mathbb{T} \setminus \{H\}. \tag{2.46b}$$

For example, for the age group 40–44 years in the breast cancer screening policy application in Chapter 3, the age group includes decision epochs from 30 to 39 so that $\beta \in \{31, 32, \ldots, 39\}$ and $\tau = 30$. Therefore, Eq. (2.43) is modified as the following,

$$\max_{\mathcal{R}} \sum_{\widehat{\varphi} \in \Phi \setminus \{\widetilde{\varphi}\}} \left\{ \widetilde{V}^0\,(\pi_x, \widetilde{\varphi}) - \widehat{V}^0\,(\pi_x, \widehat{\varphi}) \right\}, \tag{2.47a}$$

$$s.t. \quad \widetilde{V}^0\,(\pi_x, \widetilde{\varphi}) - \widehat{V}^0\,(\pi_x, \widehat{\varphi}) \geqslant 0, \text{ for } \widehat{\varphi} \in \Phi \setminus \{\widehat{\varphi}\} \tag{2.47b}$$

$$q^{\tau}_{i,\ell}(a) - \varepsilon \leqslant q^{\beta}_{i,\ell}(a) \leqslant q^{\tau}_{i,\ell}(a) + \varepsilon, \text{ for } \beta \in \{\tau+1, \tau+2, \ldots, \tau+\mathcal{L}-1\}, \tag{2.47c}$$

$$\tau = \mathcal{L}\,(\alpha(t)-1)+1, \quad \text{for } t \in \mathbb{T} \setminus \{H\}. \tag{2.47d}$$

2.6 Illustrative Example

In this section, a numerical example is presented to demonstrate the inverse algorithm in Section 2.5.

This example is a three-period and three-state POMDP problem with two transient states and one absorbing state. The underlying Markov chain is shown in Figure 2.3. From Figure 2.3, the two transient states are state 1 and 2 and the absorbing state is state 3. Also, two observation results are available for each action at each decision epoch. The observation function, O, is shown in Table 2.1a and Table 2.1b. The transition function T is shown in Table 2.1c and Table 2.1d. Note that, for simplicity, we assume that the transition matrices are equivalent given

different observation results, i.e., $T^t(a,\ell=1)=T^t(a,\ell=2)$. The end of horizon reward, q^H, is

$$q^H=\begin{bmatrix}0\\0.1\\0\end{bmatrix}.$$

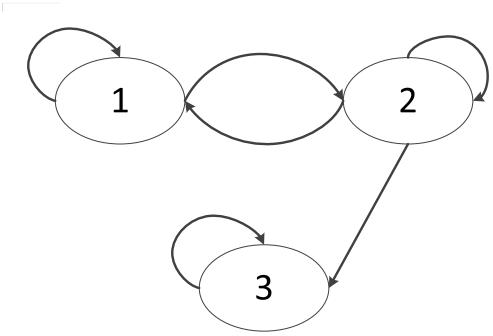


Figure 2.3: The core Markov chain of the three states inverse POMDP example.

Given all the necessary information above, the first thing to calculate is the sample path of each possible policy. The updating formulas, Eq. (2.28) and Eq. (2.30), are used to calculate the sample path for a given policy.

We use one of the possible policies, do action a_1 for all three periods, $\varphi=[a_1,a_1,a_1]$, as an example. The sample path is shown in Figure 2.4 where a node represents a belief state in which the process will be at the beginning of a particular decision epoch. The number pairs inside the nodes differentiate the nodes from each other. The first element in the pair represents the decision epoch and the second element in the pair represents the *i*-th possible outcome at the same decision epoch. For example, although the belief states associated with node (2, 2) and node (2, 4) are equivalent to each other, the two nodes are different because the paths that lead to each node are different. Node (2, 2) is the result of receiving an observation $\ell=1$ at decision

Table 2.1: The observation matrix $O^{\,t}(a^{\,t})$ given the chosen action $a^{\,t}$ and The one-step transition matrix $T^{\,t}(a,\ell)$

(a) $O^{\,t}(a^t = 1)$

$$x^{\,t} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \overbrace{\begin{array}{cc} 1 & 2 \\ \hline 1 & 0 \\ 0.9 & 0.1 \\ 1 & 0 \end{array}}^{\ell =}$$

(b) $O^{\,t}(a^t = 2)$

$$x^{\,t} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \overbrace{\begin{array}{cc} 1 & 2 \\ \hline 0.9 & 0.1 \\ 0.2 & 0.8 \\ 1 & 0 \end{array}}^{\ell =}$$

(c) $T^{\,t}(a^t = 1, \ell), \ell \in \mathcal{Z}$

$$x^t = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \overbrace{\begin{array}{ccc} 1 & 2 & 3 \\ \hline 0.9 & 0.1 & 0 \\ 0.7 & 0.2 & 0.1 \\ 0 & 0 & 1 \end{array}}^{x^{t+1}=}$$

(d) $T^{\,t}(a^t = 2, \ell), \ell \in \mathcal{Z}$

$$x^t = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \overbrace{\begin{array}{ccc} 1 & 2 & 3 \\ \hline 0.9 & 0.1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{array}}^{x^{t+1}=}$$

epoch $t = 0$ and an observation $\ell = 2$ at decision epoch $t = 1$ while node (2, 4) is the result of receiving an observation $\ell = 2$ at decision epoch $t = 0$ and an observation $\ell = 2$ at decision epoch $t = 1$. Given different paths, the nodes carries different weights when calculating the value function at decision epoch 0. Each branch connects the current belief state (origin node) and the next possible belief state (destination node) given the taken action and the possible observation. The pair associated with each branch represents the taken action and the observed result. Each branch has unique origin and destination nodes, but each node can be the origin for several branches.

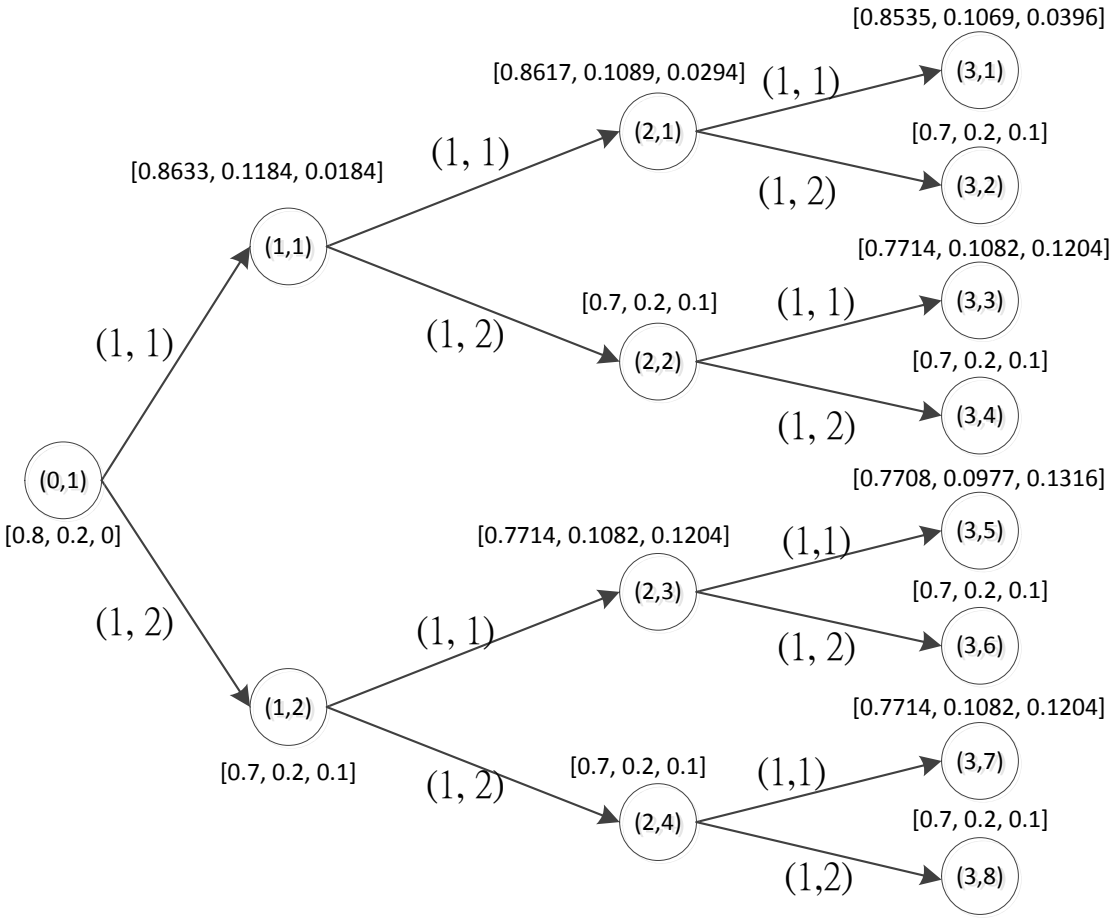


Figure 2.4: The sample path of the policy, do action one for all three periods, $\varphi = [a_1, a_1, a_1]$. A node represents a belief state in which the process will be at the beginning of a particular decision epoch. The number pairs inside the nodes differentiate the nodes from each other. The first element in the pair represents the decision epoch and the second element in the pair represents the i-th possible outcome at the same decision epoch. The number pair associated with each branch represents the taken action and the observed result.

In terms of decision variables, the value function can be calculated given the sample path. Let $V(t,k)$ denote the value function for the k-th node at decision epoch t in Figure 2.4 and $\widetilde{V}_c(t,k)$ denote the $\nu \times 1$ vector for the k-th possible belief state given the sample path, which is equivalent to the $\nu \times 1$ vector $\widetilde{V}_c^t(\pi^t)$ in Eq. (2.37) given that the k-th belief state is π^t. Using the sample path example earlier, the calculation of the value function at the node $(1,1)$ is shown below as the demonstration,

$$V(3,1) = \pi^H q^H = \begin{bmatrix} 0.8535 & 0.1069 & 0.0396 \end{bmatrix} \times \begin{bmatrix} 0 \\ 0.1 \\ 0 \end{bmatrix} = 0.0107;$$

$$\widetilde{V}_c(3,1) = \begin{bmatrix} V_1(3,1) \\ V_2(3,1) \\ V_3(3,1) \end{bmatrix} = \begin{bmatrix} 0.8535 \times 0 \\ 0.1069 \times 0.1 \\ 0.0396 \times 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0.0107 \\ 0 \end{bmatrix};$$

$$V(3,2) = \begin{bmatrix} 0.7 & 0.2 & 0.1 \end{bmatrix} \times \begin{bmatrix} 0 \\ 0.1 \\ 0 \end{bmatrix} = 0.0200;$$

$$\widetilde{V}_c(3,2) = \begin{bmatrix} V_1(3,2) \\ V_2(3,2) \\ V_3(3,2) \end{bmatrix} = \begin{bmatrix} 0.7 \times 0 \\ 0.2 \times 0.1 \\ 0.1 \times 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0.02 \\ 0 \end{bmatrix};$$

$$V(3,3) = \begin{bmatrix} 0.7714 & 0.1082 & 0.1204 \end{bmatrix} \times \begin{bmatrix} 0 \\ 0.1 \\ 0 \end{bmatrix} = 0.0108;$$

$$\widetilde{V}_c(3,3) = \begin{bmatrix} V_1(3,3) \\ V_2(3,3) \\ V_3(3,3) \end{bmatrix} = \begin{bmatrix} 0.7714 \times 0 \\ 0.1082 \times 0.1 \\ 0.1204 \times 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0.01082 \\ 0 \end{bmatrix};$$

$$V(3,4) = \begin{bmatrix} 0.7 & 0.2 & 0.1 \end{bmatrix} \times \begin{bmatrix} 0 \\ 0.1 \\ 0 \end{bmatrix} = 0.0200;$$

$$\widetilde{V}_c(3,4) = \begin{bmatrix} V_1(3,4) \\ V_2(3,4) \\ V_3(3,4) \end{bmatrix} = \begin{bmatrix} 0.7 \times 0 \\ 0.2 \times 0.1 \\ 0.1 \times 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0.02 \\ 0 \end{bmatrix};$$

$$V(2,1) = \boldsymbol{\pi}^t \sum_{\ell \in \mathbb{Z}} \left\{ \mathbb{O}^{\,t}(a,\ell) q^t(a,\ell) \right\} + \boldsymbol{\pi}^t \sum_{\ell \in \mathbb{Z}} \left\{ \mathbb{O}^{\,t}(a,\ell) T^{\,t}(a,\ell) \widetilde{V}_c^{t+1} \left(\boldsymbol{\pi}^{t+1}(a,\ell) \right) \right\}$$

$$= \begin{bmatrix} 0.8617 & 0.1089 & 0.0294 \end{bmatrix} \times \left\{ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.9 & 0 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} q^3_{1,1}(1) \\ q^3_{2,1}(1) \\ q^3_{3,1}(1) \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} q^3_{1,2}(1) \\ q^3_{2,2}(1) \\ q^3_{3,2}(1) \end{bmatrix} \right\}$$

$$+ \begin{bmatrix} 0.8617 & 0.1089 & 0.0294 \end{bmatrix} \times \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.9 & 0 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} 0.9 & 0.1 & 0 \\ 0.7 & 0.2 & 0.1 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} V_1(3,1) \\ V_2(3,1) \\ V_3(3,1) \end{bmatrix}$$

$$+ \begin{bmatrix} 0.8617 & 0.1089 & 0.0294 \end{bmatrix} \times \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 0.9 & 0.1 & 0 \\ 0.7 & 0.2 & 0.1 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} V_1(3,2) \\ V_2(3,2) \\ V_3(3,2) \end{bmatrix},$$

$$= 0.8617\, q^3_{1,1}(1) + 0.0980\, q^3_{2,1}(1) + 0.0294\, q^3_{3,1}(1)$$

$$+ 0\, q^3_{1,2}(1) + 0.0109\, q^3_{2,2}(1) + 0\, q^3_{3,2}(1) + 0.0012;$$

$$V(2,2) = \begin{bmatrix} 0.7 & 0.2 & 0.1 \end{bmatrix} \times \left\{ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.9 & 0 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} q^3_{1,1}(1) \\ q^3_{2,1}(1) \\ q^3_{3,1}(1) \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} q^3_{1,2}(1) \\ q^3_{2,2}(1) \\ q^3_{3,2}(1) \end{bmatrix} \right\}$$

$$+\begin{bmatrix}0.7 & 0.2 & 0.1\end{bmatrix}\times\begin{bmatrix}1 & 0 & 0\\ 0 & 0.9 & 0\\ 0 & 0 & 1\end{bmatrix}\times\begin{bmatrix}0.9 & 0.1 & 0\\ 0.7 & 0.2 & 0.1\\ 0 & 0 & 1\end{bmatrix}\times\begin{bmatrix}V_1(3,3)\\ V_2(3,3)\\ V_3(3,3)\end{bmatrix}$$

$$+\begin{bmatrix}0.7 & 0.2 & 0.1\end{bmatrix}\times\begin{bmatrix}0 & 0 & 0\\ 0 & 0.1 & 0\\ 0 & 0 & 0\end{bmatrix}\times\begin{bmatrix}0.9 & 0.1 & 0\\ 0.7 & 0.2 & 0.1\\ 0 & 0 & 1\end{bmatrix}\times\begin{bmatrix}V_1(3,4)\\ V_2(3,4)\\ V_3(3,4)\end{bmatrix},$$

$$= 0.7000\, q^3_{1,1}(1) + 0.1800\, q^3_{2,1}(1) + 0.1000\, q^3_{3,1}(1)$$
$$+ 0\, q^3_{1,2}(1) + 0.0200\, q^3_{2,2}(1) + 0\, q^3_{3,2}(1) + 0.0012;$$

$$V(1,1) = \begin{bmatrix}0.8633 & 0.1184 & 0.0184\end{bmatrix}\times\left\{\begin{bmatrix}1 & 0 & 0\\ 0 & 0.9 & 0\\ 0 & 0 & 1\end{bmatrix}\times\begin{bmatrix}q^2_{1,1}(1)\\ q^2_{2,1}(1)\\ q^2_{3,1}(1)\end{bmatrix}+\begin{bmatrix}0 & 0 & 0\\ 0 & 0.1 & 0\\ 0 & 0 & 0\end{bmatrix}\times\begin{bmatrix}q^2_{1,2}(1)\\ q^2_{2,2}(1)\\ q^2_{3,2}(1)\end{bmatrix}\right\}$$

$$+\begin{bmatrix}0.8633 & 0.1184 & 0.0184\end{bmatrix}\times\begin{bmatrix}1 & 0 & 0\\ 0 & 0.9 & 0\\ 0 & 0 & 1\end{bmatrix}\times\begin{bmatrix}0.9 & 0.1 & 0\\ 0.7 & 0.2 & 0.1\\ 0 & 0 & 1\end{bmatrix}\times\begin{bmatrix}V_1(2,1)\\ V_2(2,1)\\ V_3(2,1)\end{bmatrix}$$

$$+\begin{bmatrix}0.8633 & 0.1184 & 0.0184\end{bmatrix}\times\begin{bmatrix}0 & 0 & 0\\ 0 & 0.1 & 0\\ 0 & 0 & 0\end{bmatrix}\times\begin{bmatrix}0.9 & 0.1 & 0\\ 0.7 & 0.2 & 0.1\\ 0 & 0 & 1\end{bmatrix}\times\begin{bmatrix}V_1(2,2)\\ V_2(2,2)\\ V_3(2,2)\end{bmatrix},$$

$$= 0.8633\, q^2_{1,1}(1) + 0.1065\, q^2_{2,1}(1) + 0.0184\, q^2_{3,1}(1)$$
$$+ 0\, q^2_{1,2}(1) + 0.0118\, q^2_{2,2}(1) + 0\, q^2_{3,2}(1)$$
$$+ 1.3674\, q^3_{1,1}(1) + 0.0310\, q^3_{2,1}(1) + 0.0330\, q^3_{3,1}(1)$$
$$+ 0\, q^3_{1,2}(1) + 0.0034\, q^3_{2,2}(1) + 0\, q^3_{3,2}(1) + 0.00082.$$

Table 2.2 shows the coefficients of all unknowns $q^t_{i,\ell}(a)$ and the constants for the value function result associated with every possible policy.

Table 2.2: The coefficients of the unknown $q^t_{i,\ell}(a)$ for the value function given a policy.

	Policy, φ							
	φ_1	φ_2	φ_3	φ_4	φ_5	φ_6	φ_7	φ_8
$q^t_{i,\ell}(a)$	$[a_1,a_1,a_1]$	$[a_2,a_1,a_1]$	$[a_1,a_2,a_1]$	$[a_2,a_2,a_1]$	$[a_1,a_1,a_2]$	$[a_2,a_1,a_2]$	$[a_1,a_2,a_2]$	$[a_2,a_2,a_2]$
$q^1_{2,1}(1)$	0.18	0	0.18	0	0.18	0	0.18	0
$q^1_{3,1}(1)$	0	0	0	0	0	0	0	0
$q^1_{1,2}(1)$	0	0	0	0	0	0	0	0
$q^1_{2,2}(1)$	0.02	0	0.02	0	0.02	0	0.02	0
$q^1_{3,2}(1)$	0	0	0	0	0	0	0	0
$q^1_{1,1}(2)$	0	0.72	0	0.72	0	0.72	0	0.72
$q^1_{2,1}(2)$	0	0.04	0	0.04	0	0.04	0	0.04
$q^1_{3,1}(2)$	0	0	0	0	0	0	0	0
$q^1_{1,2}(2)$	0	0.08	0	0.08	0	0.08	0	0.08
$q^1_{2,2}(2)$	0	0.16	0	0.16	0	0.16	0	0.16
$q^1_{3,2}(2)$	0	0	0	0	0	0	0	0
$q^2_{1,1}(1)$	1.36980	0.71763	0	0	1.36980	0.71763	0	0
$q^2_{2,1}(1)$	0.03343	0.00797	0	0	0.03343	0.00797	0	0
$q^2_{3,1}(1)$	0.02102	0.59649	0	0	0.02102	0.59649	0	0
$q^2_{1,2}(1)$	0	0	0	0	0	0	0	0
$q^2_{2,2}(1)$	0.00371	0.00089	0	0	0.00371	0.00089	0	0
$q^2_{3,2}(1)$	0	0	0	0	0	0	0	0
$q^2_{1,1}(2)$	0	0	1.23282	0.64587	0	0	1.23282	0.64587
$q^2_{2,1}(2)$	0	0	0.00743	0.00177	0	0	0.00743	0.00177
$q^2_{3,1}(2)$	0	0	0.02102	0.59649	0	0	0.02102	0.59649
$q^2_{1,2}(2)$	0	0	0.13698	0.07176	0	0	0.13698	0.07176
$q^2_{2,2}(2)$	0	0	0.02971	0.00709	0	0	0.02971	0.00709
$q^2_{3,2}(2)$	0	0	0	0	0	0	0	0
$q^3_{1,1}(1)$	1.41160	0.68642	1.16069	0.59748	0	0	0	0
$q^3_{2,1}(1)$	0.02205	0.00975	0.01290	0.00664	0	0	0	0

Table 2.2: (Continued)

	Policy, φ							
	φ_1	φ_2	φ_3	φ_4	φ_5	φ_6	φ_7	φ_8
$q^t_{i,\ell}(a)$	$[a_1,a_1,a_1]$	$[a_2,a_1,a_1]$	$[a_1,a_2,a_1]$	$[a_2,a_2,a_1]$	$[a_1,a_1,a_2]$	$[a_2,a_1,a_2]$	$[a_1,a_2,a_2]$	$[a_2,a_2,a_2]$
$q^3_{3,1}(1)$	0.03447	0.61282	0.18968	0.71402	0	0	0	0
$q^3_{1,2}(1)$	0	0	0	0	0	0	0	0
$q^3_{2,2}(1)$	0.00245	0.00108	0.00143	0.00074	0	0	0	0
$q^3_{3,2}(1)$	0	0	0	0	0	0	0	0
$q^3_{1,1}(2)$	0	0	0	0	1.27044	0.61778	1.04462	0.53773
$q^3_{2,1}(2)$	0	0	0	0	0.00490	0.00217	0.00287	0.00148
$q^3_{3,1}(2)$	0	0	0	0	0.03447	0.61282	0.18968	0.71402
$q^3_{1,2}(2)$	0	0	0	0	0.14116	0.06864	0.11607	0.05975
$q^3_{2,2}(2)$	0	0	0	0	0.01960	0.00866	0.01146	0.00590
$q^3_{3,2}(2)$	0	0	0	0	0	0	0	0
Constant	0.00064	0.00045	0.00056	0.00039	0.00052	0.00036	0.00046	0.00032

After calculating the value function in terms of the decision variables, we can construct the inverse problem Eq. (2.42). Assume that the policy $\widetilde{\varphi} = [a_2, a_1, a_2]$ (i.e., perform action a_2 at the first and the third decision epochs and a_1 at the second decision epoch) is optimal and the upper bound and the lower bound of all decision variables are one and zero respectively. Note that the forward POMDP problem of this numerical example is designed to be a minimization problem in order to more intuitively match the breast cancer screening policy application in Chapter 3, which is a minimization problem. Therefore, the optimality condition is for every other policy $\widehat{\varphi} \in \Phi \setminus \{\widetilde{\varphi}\}$, the value function result should be greater than or at least equal to the optimal policy $\widetilde{\varphi}$,

$$\widehat{V}^0\left(\pi^0, \widehat{\varphi}\right) - \widetilde{V}^0\left(\pi^0, \widetilde{\varphi}\right) \geqslant 0, \qquad \text{for all } \pi^0 \in \Pi, \widehat{\varphi} \in \Phi \setminus \{\widetilde{\varphi}\}. \tag{2.48}$$

Hence, this numerical inverse problem is revised as the following,

$$\max_{\mathcal{R}} \sum_{\widehat{\varphi} \in \Phi \setminus \{\widetilde{\varphi}\}} \left\{ \widehat{V}^0\left(\pi^0, \widehat{\varphi}\right) - \widetilde{V}^0\left(\pi^0, \widetilde{\varphi}\right) \right\}, \tag{2.49a}$$

$$s.t. \quad \widehat{V}^0\left(\pi^0, \widehat{\varphi}\right) - \widetilde{V}^0\left(\pi^0, \widetilde{\varphi}\right) \geqslant 0, \qquad \text{for all } \widehat{\varphi} \in \Phi \setminus \{\widetilde{\varphi}\}, \tag{2.49b}$$

$$0 \leqslant q^t_{i,\ell}(a) \leqslant 1, \qquad \text{for } i \in \Omega, \ell \in \mathcal{Z}, a \in \mathcal{A}, t \in \mathbb{T} \setminus H. \tag{2.49c}$$

The coefficients of the inverse program, the right-hand-side and the optimal solution are shown in Table 2.3.

Table 2.3: The coefficients of the inverse problem where the policy $\widetilde{\varphi} = \varphi_6 = [a_2, a_1, a_2]$ is the optimal policy and the initial belief state is $[0.8, 0.2, 0]$

unknowns	Coefficients of the unknowns								Optimal
	Objective	constraints, $\widehat{\varphi} - \widetilde{\varphi}$							
$q^t_{i,\ell}(a^t)$	function	$\varphi_1 - \varphi_6$	$\varphi_2 - \varphi_6$	$\varphi_3 - \varphi_6$	$\varphi_4 - \varphi_6$	$\varphi_5 - \varphi_6$	$\varphi_7 - \varphi_6$	$\varphi_8 - \varphi_6$	Solution
$q^1_{1,1}(1)$	3.2	0.8	0	0.8	0	0.8	0.8	0	1
$q^1_{2,1}(1)$	0.72	0.18	0	0.18	0	0.18	0.18	0	1
$q^1_{3,1}(1)$	0	0	0	0	0	0	0	0	0
$q^1_{1,2}(1)$	0	0	0	0	0	0	0	0	0
$q^1_{2,2}(1)$	0.08	0.02	0	0.02	0	0.02	0.02	0	1
$q^1_{3,2}(1)$	0	0	0	0	0	0	0	0	0
$q^1_{1,1}(2)$	-2.88	-0.72	0	-0.72	0	-0.72	-0.72	0	0
$q^1_{2,1}(2)$	-0.16	-0.04	0	-0.04	0	-0.04	-0.04	0	0
$q^1_{3,1}(2)$	0	0	0	0	0	0	0	0	0
$q^1_{1,2}(2)$	-0.32	-0.08	0	-0.08	0	-0.08	-0.08	0	0
$q^1_{2,2}(2)$	-0.64	-0.16	0	-0.16	0	-0.16	-0.16	0	0
$q^1_{3,2}(2)$	0	0	0	0	0	0	0	0	0
$q^2_{1,1}(1)$	-1.5662	0.652164	0	-0.71763	-0.71763	0.652164	-0.71763	-0.71763	0
$q^2_{2,1}(1)$	0.019015	0.025455	0	-0.00797	-0.00797	0.025455	-0.00797	-0.00797	1
$q^2_{3,1}(1)$	-3.53691	-0.57547	0	-0.59649	-0.59649	-0.57547	-0.59649	-0.59649	0
$q^2_{1,2}(1)$	0	0	0	0	0	0	0	0	0
$q^2_{2,2}(1)$	0.002113	0.002828	0	-0.00089	-0.00089	0.002828	-0.00089	-0.00089	1
$q^2_{3,2}(1)$	0	0	0	0	0	0	0	0	0
$q^2_{1,1}(2)$	3.757369	0	0	1.232816	0.645868	0	1.232816	0.645868	1
$q^2_{2,1}(2)$	0.018401	0	0	0.007429	0.001772	0	0.007429	0.001772	1
$q^2_{3,1}(2)$	1.235023	0	0	0.02102	0.596491	0	0.02102	0.596491	1
$q^2_{1,2}(2)$	0.417485	0	0	0.13698	0.071763	0	0.13698	0.071763	1
$q^2_{2,2}(2)$	0.073604	0	0	0.029714	0.007088	0	0.029714	0.007088	1
$q^2_{3,2}(2)$	0	0	0	0	0	0	0	0	0
$q^3_{1,1}(1)$	3.856195	1.4116	0.68642	1.160693	0.597482	0	0	0	1
$q^3_{2,1}(1)$	0.051334	0.022051	0.009747	0.012897	0.006639	0	0	0	1
$q^3_{3,1}(1)$	1.550993	0.034472	0.612821	0.189681	0.714019	0	0	0	1
$q^3_{1,2}(1)$	0	0	0	0	0	0	0	0	0

Table 2.3: (Continued)

unknowns	Coefficients of the unknowns								Optimal
	Objective	constraints, $\widehat{\varphi} - \widetilde{\varphi}$							
$q^t_{i,\ell}(a^t)$	function	$\varphi_1 - \varphi_6$	$\varphi_2 - \varphi_6$	$\varphi_3 - \varphi_6$	$\varphi_4 - \varphi_6$	$\varphi_5 - \varphi_6$	$\varphi_7 - \varphi_6$	$\varphi_8 - \varphi_6$	Solution
$q^3_{2,2}(1)$	0.005704	0.00245	0.001083	0.001433	0.000738	0	0	0	1
$q^3_{3,2}(1)$	0	0	0	0	0	0	0	0	0
$q^3_{1,1}(2)$	-1.47165	-0.61778	-0.61778	-0.61778	-0.61778	0.652663	0.426846	-0.08004	0
$q^3_{2,1}(2)$	-0.00592	-0.00217	-0.00217	-0.00217	-0.00217	0.002734	0.0007	-0.00069	0
$q^3_{3,1}(2)$	-3.35158	-0.61282	-0.61282	-0.61282	-0.61282	-0.57835	-0.42314	0.101198	0
$q^3_{1,2}(2)$	-0.16352	-0.06864	-0.06864	-0.06864	-0.06864	0.072518	0.047427	-0.00889	0
$q^3_{2,2}(2)$	-0.02368	-0.00866	-0.00866	-0.00866	-0.00866	0.010937	0.002799	-0.00276	0
$q^3_{3,2}(2)$	0	0	0	0	0	0	0	0	0
RHS		0.00374	0.00176	0.00242	0.00088	0.00154	0.0007	-0.00056	

Table 2.4: The optimal solution of the numerical example with the age group concept.

	Optimal Solution		Optimal Solution
$q_{1,1}(1)$	0.97394	$q_{1,1}(2)$	1
$q_{2,1}(1)$	1	$q_{2,1}(2)$	0
$q_{3,1}(1)$	0.86042	$q_{3,1}(2)$	0.84905
$q_{1,2}(1)$	1	$q_{1,2}(2)$	1
$q_{2,2}(1)$	1	$q_{2,2}(2)$	0
$q_{3,2}(1)$	1	$q_{3,2}(2)$	1

From the example, we can see that all unknowns are equal to zero or one. From the perspective of simply imputing a set of rewards that makes a particular policy the optimal policy, the optimal solution that Eq. (2.49) returns does return the smallest value function of the policy, perform a_2 at the first and the third decision epoch and a_1 at the second decision epoch $\varphi_6 = [a_2, a_1, a_2]$, among all possible policies. However, in a real-life application such as the breast cancer screening policy application in Chapter 3, the optimal solution obtained from Eq. (2.49) does not provide an insightful answer to a disease screening problem. Therefore, we propose to apply the age group α concept to the inverse problem, which adds the constraint that the unknowns $q^t_{i,\ell}(a)$ are equal to each other within the same age group,

$$q_{i,\ell}(a) = q^1_{i,\ell}(a) = q^2_{i,\ell}(a) = q^3_{i,\ell}(a), i \in \Omega, \ell \in \mathcal{Z}, a \in \mathcal{A}, \tag{2.50}$$

in this numerical example assuming only one age group. Therefore, the decision variable $q_{i,\ell}(a)$ is a weighted average cost in an age group. Table 2.4 shows the optimal solution after applying the age group concept.

Though the equality constraint brings more operationally feasible solutions into the inverse problem, there are trade-offs associated with this approach. The equality constraints also

make the inverse problem for some policies infeasible — for example, the breast cancer screening policy problem in Chapter 3.

With regard to the computational effort, experiments with different time horizon lengths are solved. The numerical example above serves as the basic example. The tested time horizon length starts from three decision epochs. A time horizon length forms an experiment. For each time horizon length, all possible time-dependent policies are tested. In other words, after calculating the sample paths for all time-dependent policies, each time-dependent policy is assumed to be the optimal policy. Experiments are solved with on a 64-bit Intel Core2 Quad 2.40 GHz CPU with 8MB cache. Table 2.5 shows the computer time for each experiment. The first column is the number of decision epochs the experiment solves. The computer times for calculating the sample path and the whole program are shown in the second column and the third column respectively.

Table 2.5: Computer time for solving different time horizon length.

	Computer Time (seconds)	
Time horizon length	Sample path	The whole program
3	0.04865	0.07305
5	1.20982	1.32533
7	47.55369	48.70477
9	2565.06975	2590.58630

From Table 2.5, calculation of the sample paths consumes most of the computational effort within an experiment, even when the whole program solves all inverse problems for each possible time-dependent policy.

2.7 Conclusion

In order to solve an inverse problem, a comprehensive understanding of the associated forward problem is necessary. Hence, we illustrate a general discrete-time POMDP with finite states and a finite time horizon. Moreover, a POMDP designed specifically for disease screening is introduced. Given the switch of the order of taking an observation and making a state transition in the core process, the differences, such as the definitions of the transitions and the immediate rewards and the formulation of the value function, are discussed. The updating formulas, the most important difference between a POMDP that is normally introduced in the literature and a disease screening POMDP, are described in detail.

We develop an solution method to solve the inverse POMDP problem with time-dependent policies. The main idea of our proposed method is this: we seek to maximize the summation of the differences of the value function results so that the value function of the specified policy is as far away as possible from the value function of other policies at time 0. This optimization is subject to constraints requiring that the value function result of the specified policy should be greater than or at least equal to the value function of all other possible time-dependent policies at decision epoch 0. The belief state space forms a part of the constraint set. In this maximization problem, the decision variables are the initial belief state $\pi^0 \in \Pi$ and the rewards $q^t_{i,\ell}(a)$ for $i \in \Omega$, $a \in \mathcal{A}$, $\ell \in \mathcal{Z}$ and $t \in \mathbb{T} \setminus \{H\}$. Therefore, the maximization problem is a nonlinear problem given the definition of the value function Eq. (2.33). To solve the nonlinear program, we convert the inverse problem formulation from a nonlinear problem into a linear problem by discretizing the belief state space Π.

Our proposed method also considers the degeneracy problem that one typically encounters while solving an inverse problem. We propose to maximize the summation of the differences of the value function results so that the value function of the specified policy is as far away as

possible from the value function of other possible actions.

To search for an operationally feasible answer of a real-life application, we propose to add an equality constraint to the rewards associated with the same observation, action, and state, but different decision epochs as shown in Eq. (2.45). However, this equality constraint is a very strong assumption which may result in infeasibility of the inverse problem. To resolve the infeasibility problem, a robust method is applied to the equality constraint to allow some tolerance in the equality constraint shown in Eq. (2.46).

A numerical example is provided to show the inverse problem calculation. We show the optimal solution of the inverse problem without the equality constraint. The solutions are either zero or one, which does not provide an insightful solution to a disease screening problem. Therefore, the equality constraint and the robust method are applied and the result is provided to demonstrate the difference of applying the equality constraint. Also, the computational effort of solving the small example is presented in Table 2.5.

Chapter 3

The Inverse POMDP Model of Breast Cancer Screening Policies

3.1 Introduction

Breast cancer is a disease caused by the formation of malignant cancer cells in the tissues of the breast. Breast cancer is the most common type of cancer among women in the United States other than skin cancer, accounting for nearly 1 in 3 cancers diagnosed in American women [43]. The American Cancer Society estimates that 232,340 new cases of female breast cancer will be diagnosed, and 39,620 deaths will occur in the United States in 2013 [44]. It is estimated that about 12% of women born today will be diagnosed with breast cancer at some point during their lifetimes [25]. In other words, on average, 1 out of 8 women will be diagnosed with breast cancer during her lifetime. Breast cancer ranks second among top causes of cancer death in women after lung cancer [44].

Although breast cancer is a highly fatal disease, improvements in breast cancer treatment and early detection have contributed greatly to the decline in breast cancer mortality. Breast

cancer death rates decreased 2.2% per year between 1990 and 2007, especially for younger women [23]. The five-year relative survival for women diagnosed with localized breast cancer (i.e., cancer that has not spread to lymph nodes or other locations outside the breast) is 99%; if the cancer has spread to nearby lymph nodes (i.e., the cancer is in the regional stage), or distant lymph nodes or organs (i.e., the cancer is in the distant stage), then the survival rate falls to 84% and 23%, respectively [44].

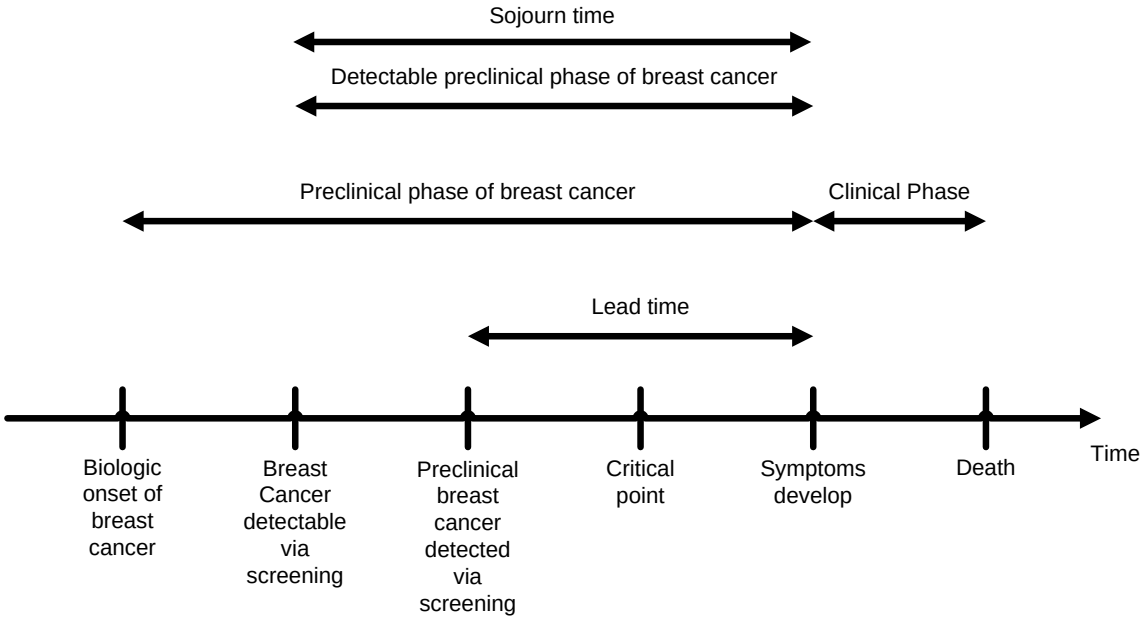


Figure 3.1: Diagram of the natural history of breast cancer. [20]

The timeline in Figure 3.1 [20] shows the natural progression of breast cancer from its onset until death. The progression from biologic onset of breast cancer to death is divided into preclinical and clinical phases. The detectable preclinical phase of breast cancer is the period during which screening tests are applied to detect a condition early in its natural history, before

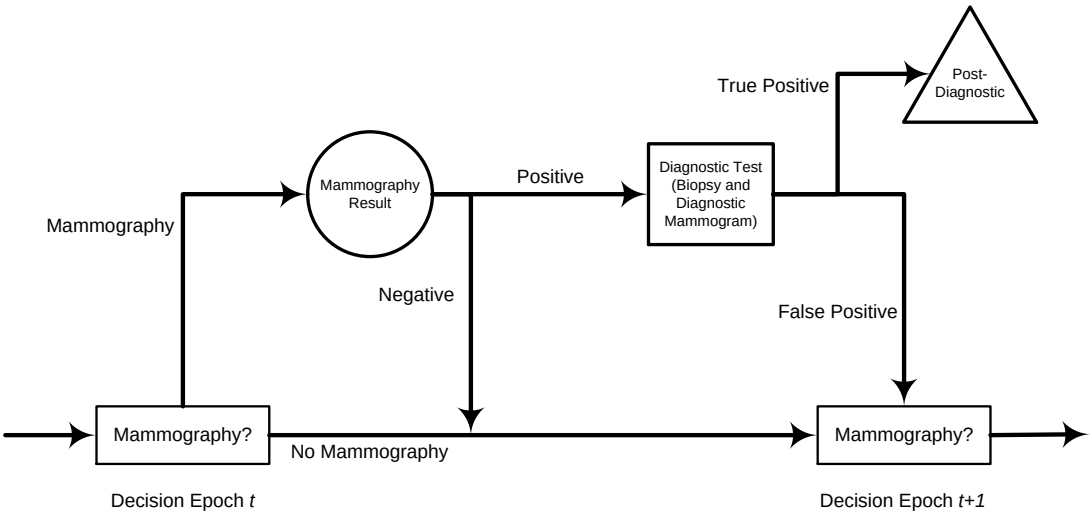


Figure 3.2: The timeline of the breast cancer screening decision process

the onset of symptoms [20]. In the early stage, i.e., in the preclinical phase, breast cancer normally does not produce any symptoms. However, medical studies show that treatment is more effective and a cure is more likely during the preclinical phase. Therefore, a series of regular screening tests over a woman's lifetime, known as a screening policy, is recommended to detect breast cancer before symptoms develop. Mammography, a low-dose X-ray procedure that allows visualization of the internal structure of the breast, is the most recommended regular screening test with high accuracy for early detection of breast cancer, where the accuracy of the mammography is measured by the sensitivity and the specificity of the test. Sensitivity is the probability that the test will detect breast cancer when the patient has the disease (i.e., yield a true positive result); and specificity is the probability that the test will give a negative result when the patient does not have the disease (i.e., yield a true negative result). Also, the accuracy is a function of the patient age. Please see Table 4.5 for the sensitivity and specificity of the mammography for different patient ages.

Figure 3.2 illustrates all possible events between two decision epochs. At decision epoch

t, a decision is made regarding whether or not to perform a mammogram. If the decision is not to perform a mammogram, then no event will occur until the next decision epoch. If the decision is to take a mammogram, then the result of the mammogram will be either positive or negative. Given the result is negative, no other event will occur until the next decision epoch. On the other hand, if the mammogram is positive, then because the mammogram is not a perfect test, a follow-up test, such as a biopsy, will take place to reveal if the first mammogram is a true positive or a false positive. If the follow-up test shows the screening mammogram is a false positive, then the next possible event will be the decision of whether to take a screening mammogram at the next decision epoch. If the follow-up test shows the screening result at the current decision epoch is a true positive, then the screening mammogram decision process will end with the patient moving to the postdiagnostic phase, where medical intervention may occur to remove the cancer tumor.

Different screening policies are recommended by different organizations. For example, the American Cancer Society recommends annual screening for an average risk woman beginning at age 40 and continuing for as long as a woman is in good health [44], while the US Preventive Services Task Force recommends biennial mammograms for women from age 50 to age 74 [50]. Table 3.1 summarizes US and international breast cancer screening guidelines. Policies specify the starting age and the time interval between mammograms. In this way, a policy is relatively easy to understand and follow because of the time dependency. However, different recommendations confuse the general public. What does a screening policy say about a woman's health with regard to preventing breast cancer? From the perspective of economic analysis, we are naturally led to ask the following question: Are there some *imputed* rewards or costs for a woman associated with following a particular policy when a doctor recommends her to do so?

This chapter aims to identify the rewards for an average risk woman to follow a given

Table 3.1: Summary of US and international mammography screening recommendations

	Recommendation	Agency
US	Annually for women $\geqslant$ 40 y	American Cancer Society [44]
		American Medical Association [2]
		American College of Obstetricians and Gynecologists [36]
		American College of Radiology [28]
	Every 1–2 y for women $\geqslant$ 40 y	National Cancer Institute [24]
	Biennially for women 50–74 y	US Preventive Services Task Force (after 2009) [50]
International	Every 2–3 y for women 50–74 y	Canadian Task Force on Preventive Health Care [49]
	Every 3 y for women $\geqslant$ 50 y	NHS Breast Screening Programme in UK [35]
	Biennially for women 50–69 y	BreastScreen Australia [8]

breast cancer screening policy. We develop a model for breast cancer applied to the general population and investigate the rewards that can be imputed (attributed) to a designated breast cancer screening policy that is currently being studied, where the imputed reward is expressed as the woman's lifetime conditional probability of death caused by breast cancer, given that her cancer is detected in a particular stage at a particular age. This model searches for the imputed rewards for a screening policy such that a chosen screening policy is better than all other policies. To this end, we formulate a POMDP for breast cancer screening and we apply the inverse algorithm presented in Section 2.5. Owing to the nature of breast cancer, a woman's health status is not directly observable; nevertheless, information about her health status can be gathered through mammography with a certain level of error.

By applying the inverse algorithm to the forward breast cancer POMDP model, we can learn what the reward (a woman's lifetime breast cancer mortality probability of being detected

in a particular cancer state) should be so that a designated time-dependent policy can minimize a woman's lifetime breast cancer mortality probability. Furthermore, we can use the reward structure to compare different time-dependent policies to enable the patient to select a time-dependent policy for herself that is also easier to follow.

The remainder of this chapter is organized as follows. Section 3.2 reviews the relevant literature on MDP and POMDP models for the design and analysis of breast cancer screening policies. Inverse problems in the healthcare area are also discussed in Section 3.2. Our POMDP model is presented in Section 3.3. The formulation of the inverse problem is described in Section 3.4.

3.2 Review of Previous Work

We start this section by discussing previous work on the breast cancer screening problem. Also, some inverse problems in the healthcare area are reviewed in the second part of this section. At the end of this section, we discuss the difference between the reviewed work and our study.

Models for Breast Cancer Screening

Maillart et al. [30] focus on the breast cancer screening problem for women before and after menopause, because the natural progression of breast cancer is less aggressive in older women, but the result of mammography is more accurate for older women. The authors aim to answer the question of whether the screening guideline should be different for women before and after menopause. Also, the authors argue that screening policies prescribing different screening intervals should depend on the patient's age. They model the problem as a partially observable discrete-time finite-horizon Markov chain and evaluate 1,223 policies with different starting ages, ending ages, and screening intervals. By solving the problem with the policy evaluation

method, they consider only time-dependent policies, which are easier for the patient to follow compared with the state-dependent policies based on solving a POMDP exactly. Lifetime breast cancer mortality probability is used to assess the policy efficiency to avoid the "patient specific" problem that arises when using the quality-adjusted life-years (QALYs) because computing QALYs requires an individual's utility values, which are unique to each person. Hence, the model searches for the policy which minimizes the lifetime breast cancer mortality probability. The authors report a frontier of efficient policies that provide lower values of the lifetime breast cancer mortality probability and the lifetime expected number of mammograms. Also, Maillart et al. demonstrate the robustness of the resulting efficient frontier in the sense that policies close to the frontier are nearly Pareto optimal.

Ayer et al. [6] formulate the breast cancer screening problem as a discrete-time finite-horizon POMDP problem. The work aims to build a personalized screening policy given a patient's health status. They argue that the screening policy should consider not only a patient's age but also other factors, such as family history, breast density, body mass index (BMI), alcohol consumption, and extent of breastfeeding as well as ages at menarche, menopause, and first birth. The number of QALYs is used as the reward in the authors' POMDP model so that the goal of the model is to maximize the patient's QALYs with regard to breast cancer. In addition to the decision of whether to have a mammogram, the authors also consider the possibility that the patient detects breast cancer symptoms. Smallwood and Sondik's one-pass algorithm [42, 45] is used to solve the POMDP. Also, Ayer et al. proved that the one-pass algorithm is applicable even if the order of taking the observation and making the state transition are switched so that each event in the progression of breast cancer happens after the corresponding mammogram result is revealed.

Inverse Problems in Healthcare

Erkin et al. [17] constructed an inverse MDP approach to reveal patients' preferences using the living-donor liver transplant problem as an example. A discounted, infinite-horizon and finite-state MDP can be constructed as an LP problem [38] that can be readily solved by existing LP software systems. The solution to this LP problem is a stationary deterministic state-dependent policy, i.e., the policy is a mapping from the state space to the action space. Every action in the action space is a feasible solution of the LP problem. Hence, in order to infer the patient's preference, they apply the inverse linear programming algorithm of Ahuja and Orlin [1] as discussed in Section 2.2 to such an MDP. The proposed method seeks the new coefficients (the patient's preference) so that a particular action (a patient's behavior) is optimal by inverting the LP problem under the L_1 norm defined by Eq. (2.1). In other words, the proposed inverse algorithm is also a LP minimization problem whose objective function is the L_1 norm of the difference between the new coefficients and the original coefficients.

Lee and Zenios [29] utilize the shadow price concept to analyze the patterns of health-care demand, spending, and disparity, where the shadow price represents the society's implied willingness-to-pay (WTP) for medicine measured in monetary value per QALY. They focus on the problem of the allocation of renal dialysis (RD) resources for the end-stage renal disease (ESRD) population. The shadow price is the value of the Lagrange multiplier at optimality and indicates the improvement in the objective value that an additional unit of the constrained resource will bring about [27]. The RD problem is modeled as a constrained MDP in which the objective is to maximize the societal QALYs subject to a constraint on total spending. Therefore, the shadow price is the Lagrange multiplier, which brings the constraint to the objective function so that the problem can be modeled as an MDP problem without constraints. The shadow price is found when the QALYs and total spending match the currently observed de-

cision maker's behavior. Normally, the observed behavior does not match the theoretical result from solving the MDP problem, i.e., the observed behavior is suboptimal for the MDP problem. To resolve the suboptimality issue, they propose to apply the welfare concept of willingness-to-pay and willingness-to-accept (WTA) parity. The variable WTA measures the amount of money society demands for accepting an increased risk of death. The two should theoretically equal each other, which is known as WTP-WTA parity. When the parity is achieved, the shadow price is then found. The relationship between WTP and WTA provides the direction while searching for the shadow price. Hence, the problem can be solved more efficiently. With this proposed method, they found that the implied WTP for ESRD patient is $158,000 per QALY, and they show that the implied WTP varies according to race and age.

Distinctive Features of the Current Work

The model developed in this chapter differs from the above-referenced work in that we have developed an inverse POMDP model for imputing rewards to time-dependent breast cancer screening policies with nonstationary immediate rewards. We build a forward POMDP for breast cancer screening policy and apply the inverse algorithm of Section 2.5 to the forward POMDP for that policy, where the formulation of the inverse POMDP incorporates some constraints that are specific to the designated breast cancer screening policy. Our model assumes the designated policy is the best policy, and it imputes the lifetime breast cancer mortality probability associated with being detected in a particular stage of breast cancer at a particular age.

3.3 POMDP Model for Breast Cancer Screening Policies

In order to apply the inverse algorithm described in Section 2.5, a forward POMDP setup is necessary. In this section, we detail the formulation of a forward POMDP model for constructing breast cancer screening policy.

3.3.1 The Core Process

We model a patient's natural health progression by using a discrete-time Markov chain with five different states for a woman between the ages of 25 and 100:

- *No Breast Cancer* (state NBC);
- *Preclinical* (state P);
- *Clinical* (state C);
- *Non–Breast Cancer Induced Death* or *Death for Other Causes than Breast Cancer* (state DOBC); and
- *Postdiagnostic* (state PD).

Data from the Surveillance, Epidemiology, and End Results (SEER) database shows that women younger than 25 years old have very low age-adjusted breast cancer incidence rates per 100,000 women per year in the general population — namely, a rate of 0.2 newly diagnosed breast cancers for those aged 15–19 years and 1.6 newly diagnosed breast cancers for those aged 20–24 years [23, 52]. Another medical study found that no cancer was detected by mammography for women aged 18–24 years [52]. Hence, our POMDP model starts for an average risk woman at age 25.

Data from the National Center for Health Statistics shows that the average female life expectancy at birth is 81.1 years [31]. Also, a total of 44,202 females out of approximately 157 million females in the United States are 100 and above, which is less than 0.03% of the US female population [22]. The life expectancy combined with the extremely low percentage of women over 100 years old justifies using the upper limit of 100 for the age of women in our Markov chain model.

This Markov chain model only describes the progression before a patient is diagnosed with breast cancer. As shown in Figure 3.1, a woman in the preclinical breast cancer state has no known signs or symptoms of breast cancer, but the cancer is detectable by the screening test; on the other hand, a woman in the clinical breast cancer state shows signs or symptoms of breast cancer, such as lumps, swelling, and skin changes. We define our breast cancer states according to the natural progression of breast cancer shown in Figure 3.1 and the available data sets. Also, the sojourn time is the duration of the time period in which a disease case is occult but detectable by screening — i.e., the interval in Figure 3.1 with the label "Detectable preclinical phase of breast cancer." [13]. Hence, the states P and C are chosen as a part of the state space because of their widespread use as classification terms in the medical literature and medical studies. The two states DOBC and PD are fully observable and absorbing states. Following the system of notation set up in Chapter 2, we let Ω denote the core Markov chain state space, while Ω' denotes the subset of Ω which is composed of the transient states in Ω, and t denotes the current age of an average risk woman in whom we are interested. Hence, the whole core Markov Chain state space Ω can be defined as $\Omega = \{i : i = \text{NBC, P, C, DOBC, PD}\}$; the subset Ω' can be defined as $\Omega' = \{i : i = \text{NBC, P, C}\}$. Figure 3.3 shows the Markov chain diagram.

We assume that a patient cannot regress to a previous (better) health state without medical intervention. Hence, as shown in Figure 3.3, a patient in the state NBC can make any of the

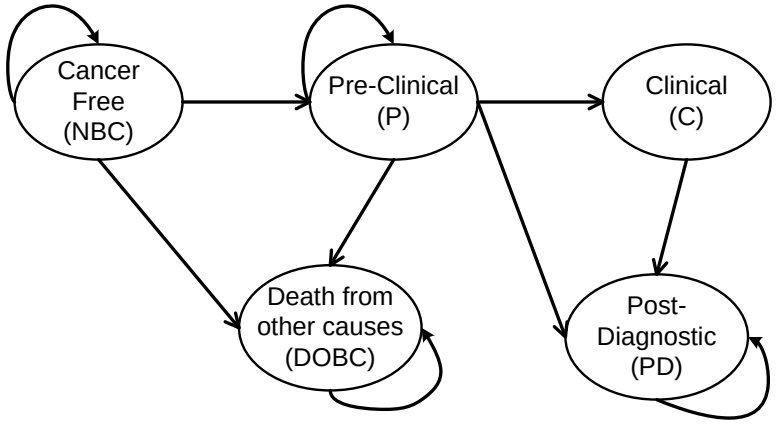


Figure 3.3: The core Markov chain representing the natural progression of breast cancer.

following state transitions: (a) move to states P or DOBC; or (b) remain in the state NBC. The state transitions of a patient in state P can be one of the following: (*i*) move to either state C or state DOBC; (*ii*) remain in state P; or (*iii*) move to state PD if the patient has an abnormal mammogram result and is correctly diagnosed with breast cancer (the arrow labelled "true positive" in Figure 3.2). The patient moves to the state PD by undergoing postdiagnostic treatment. For a patient who moves into the state C, the patient will make a transition into state PD with probability one.

3.3.2 Structure of the POMDP

Our POMDP for breast cancer screening is different from a conventional POMDP because at each decision epoch t, the observation z^t is taken before the core process makes a transition from the current state x^t to the next state x^{t+1} as discussed in Section 2.4.

Decision Epochs

Let H denote the total number of the decision epochs. The time between two decision epochs is six months, i.e., there are total 150 decision epochs, so that $H = 150$. Six months is used, instead of a year, as the interval between two decision epochs to allow a patient who might develop symptoms to receive a diagnosis before the next scheduled screening. In this way, a patient is able to leave the system before the next screening given that the interval between two screening tests for most screening policies is greater than six months.

Actions

The action set $\mathcal{A}$ includes two actions available to the patient at each decision epoch: to wait and not to screen (NS), or to take a mammographic screening test (S).

Observations

As mentioned in Section 3.1, mammography is an accurate screening test whose sensitivity and specificity are a function of age, i.e., the younger the patient is, the lower the sensitivity and the specificity are. However, mammography is not a perfect test, which means that the core Markov-chain process depicted in Figure 3.3 is only partially observable. Hence, the state NBC and the P state are partially observable, and we assume that the C state is fully observable given the definition discussed previously in Section 3.3.1. Note, this POMDP focuses on the breast cancer screening policy, i.e., the square labelled "mammography?" in Figure 3.2. Hence, the diagnosis test result is not included in the observations. For a patient in the states NBC or P, there is some probability that mammography reports a false result given that the action taken is to screen. A false result could be: (a) a false positive for a patient who is in the state NBC but receives an abnormal mammogram; or (b) a false negative result for a patient who is in the

state P but receives a normal mammogram.

A false positive result can have a devastating psychological impact on a patient and can actually have a negative effect on her total QALYs [4]. On the other hand, a false negative result might delay the patient in seeking treatment options. In other words, a patient who is in the state P and chooses to do the screening test might receive a normal mammogram and believe she does not have breast cancer. Hence, this patient would not take any action until the next screening or until breast cancer symptoms develop. To reveal if an abnormal mammogram is a true positive or a false positive, we assume that a "perfect" diagnostic test (e.g., a diagnostic mammogram or a biopsy test with sensitivity and specificity both equal to 1) is performed right after the screening mammogram result is learned at the same decision epoch t. Note that the result of the diagnostic test result is not a part of the observation result z^t in this POMDP.

At each decision epoch, two possible observations, normal (N) or abnormal (Ab), are associated with the chosen action. At decision epoch t, the probabilities of a possible observation is a conditional probability given the action and the current state at each decision epoch,

$$\begin{aligned} O^t_{i,\ell}(a) = \Pr\left\{ z^{\,t} = \ell \,\middle|\, x^{\,t} = i \text{ and the chosen action is } a \right\}, \\ i \in \Omega, \ell \in \{\text{N, Ab}\}, \text{ and } t \in \mathbb{T} \setminus \{H\}. \end{aligned} \tag{3.1}$$

The observation result $z^{\,t} \in \{\text{N, Ab}\}$ for state NBC, P and C is the combination of the mammogram result and the self-detection result. When the action is not to screen, $a =$NS, the observation result, $O^t_{i,\ell}(\text{NS})$ for $i \in \Omega'$ and $\ell \in \{\text{N, Ab}\}$, represents the self-detection result. On the other hand, when the action is to take a mammogram, $a =$S, the observation, $O^t_{i,\ell}(\text{S})$ for $i \in \Omega'$ and $\ell \in \{\text{N, Ab}\}$ represents the result from the mammogram.

The state DOBC is special because a woman who is dead will yield neither of the two observations, normal or abnormal. In order to accurately describe the status of a woman en-

tering the state DOBC, the probabilities of both observations are assigned to be zero, i.e., $O^t_{\text{DOBC},\ell}(a) = 0$ where $\ell \in \{\text{N, AB}\}$ and $a \in \mathcal{A}$. Moreover, an observation, Death, is added to the observation space as discussed earlier in Section 2.4 to capture that the process will stay in state DOBC for the remaining decision epochs. Hence, the conditional probability that the process is in state DOBC and receives the observation, Death, is equal to one, i.e., $O^t_{\text{DOBC, Death}}(a) = 1$ where $a \in \mathcal{A}$. Similar to the state DOBC, a patient in state PD does not yield the observation, normal (ℓ =N) or abnormal (ℓ =Ab), i.e., $O^t_{\text{PD},\ell}(a) = 0$ where $\ell \in \{\text{N, AB}\}$ and $a \in \mathcal{A}$. Also, the observation, PD, is added to the observation space

$$\mathcal{Z} = \{\text{N, Ab, Death, PD}\} \text{ and } \zeta = 4; \tag{3.2}$$

and the conditional probability that the process is in state PD and receives the observation, PD, is equal to one, i.e., $O^t_{\text{PD, PD}}(a) = 1$ where $a \in \mathcal{A}$. For states NBC, P, and C, the conditional probability of receiving the observation, Death or PD, is equal to zero, i.e., $O^t_{i,\ell}(a) = 0$ where $i \in \Omega'$, $\ell \in \{\text{Death, PD}\}$, and $a \in \mathcal{A}$.

For a woman of age t, if the chosen action to follow is to wait and do nothing (a^t =NS), then provided that the woman's current health state is NBC or P so that $x^t \in \{\text{NBC, P}\}$, the observation must be normal, z^t =N, according to the definition of the states NBC and P. In other words, the probability of a normal observation for a woman in either the state NBC or P is one as a preclinical breast cancer cannot be detected without a mammogram. On the other hand, if a woman of age t is in the state C, then regardless of which the action is, a^t =NS or a^t =S, we have

$$O^t_{\text{C, Ab}}(a^t) = \Pr\left\{ z^t = \text{Ab} \middle| x^t = \text{C and action } a^t \text{ is taken} \right\} = 1 \text{ for all } a^t \in \mathcal{A}, t \in \mathbb{T} \setminus \{H\}. \tag{3.3}$$

A "perfect" diagnostic test is assumed to be performed at the same decision epoch t to confirm the patient is in the state C.

If the chosen action to follow is to screen, a^t =S, then a woman at age t might receive an abnormal mammogram with probability $O^t_{i,\text{Ab}}(\text{S})$, if $i \in \Omega'$ so that she is in one of the transient states, NBC, P or C. Table 3.2 summarizes the conditional probability of the observations in the matrix form.

Table 3.2: The observation matrix $O^t_{i,\ell}(a^t)$ for a woman at age t given the chosen action

(a) *No Screening :* a^t =NS

$$x^t = \begin{array}{c} \\ \text{NBC} \\ \text{P} \\ \text{C} \\ \text{DOBC} \\ \text{PD} \end{array} \overbrace{\begin{array}{c} \begin{array}{cccc} \text{N} & \text{Ab} & \text{Death} & \text{PD} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}}^{\ell =}$$

(b) *Screening :* a^t =S

$$x^t = \begin{array}{c} \\ \text{NBC} \\ \text{P} \\ \text{C} \\ \text{DOBC} \\ \text{PD} \end{array} \overbrace{\begin{array}{c} \begin{array}{cccc} \text{N} & \text{Ab} & \text{Death} & \text{PD} \end{array} \\ \begin{bmatrix} 1-O^t_{\text{NBC,Ab}} & O^t_{\text{NBC,Ab}} & 0 & 0 \\ 1-O^t_{\text{P,Ab}} & O^t_{\text{P,Ab}} & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}}^{\ell =}$$

Transition Matrix

At each decision epoch $t \in \mathbb{T} \setminus \{H\}$, let $T_{i,j}^{\ t}(a^t, z^t)$ denote the $(i,j)^{th}$ element of the one-step transition probability matrix, $T^{\ t}$, for a woman in state x^t at age t given action $a^t \in \mathcal{A}$ and the observation result $z^t \in \mathcal{Z}$. As we mentioned previously, the transition probability is a conditional probability given the action a and the observation result $z^{\ t}$. However, the mammography does not change the natural progression of breast cancer, which means that if the observation result is normal, then the transition probability matrix of both available actions, to wait and not to screen (NS) or to take a screening test (S), are identical to each other, i.e., $T^{\ t}(\text{NS, N}) \equiv T^{\ t}(\text{S, N})$.

A patient of age t who is in the state NBC can move to either the state P or the state DOBC with probability $T_{\text{NBC},j}^{t}(a,\text{N})$, $j \in \{\text{P, DOBC}\}$; or she can stay in the state NBC with probability $T_{\text{NBC, NBC}}^{t}(a,\text{N})$.

If a patient of age t years is in state P with a normal observation, then she is not aware of her breast cancer without a mammogram by the definition of the state P. Given the assumption that a patient cannot regress to a previous state of improved health without a medical intervention, a patient of age t years in state P can only transfer into states C or DOBC with probabilities $T_{\text{P},j}^{t}(a,\text{N})$, $j \in \{\text{C, DOBC}\}$; or she can stay in the state P with probability $T_{\text{P, P}}^{t}(a,\text{N})$. Note that, for a patient who is in the state NBC or P, we assume that the probabilities of moving into the state DOBC is the same, i.e., $T_{\text{NBC, DOBC}}^{t}(a,\text{N}) = T_{\text{P, DOBC}}^{t}(a,\text{N})$, because, for a patient who is in the state P, she will not die from breast cancer without progressing to state C in a one-step transition so that the probability of dying from other causes while she is in the state P is equal to the probability of dying from other causes when she is in the state NBC. Therefore, we use T_{DOBC}^{t} to represent the transition probability going into the state DOBC from the state NBC or P.

A patient of age t who is in the state C will move into the state PD with probability one,

i.e., $T^{t}_{\text{C, PD}}(a,\text{N}) = 1$ for all $a \in \mathcal{A}$ and $t \in \{0,1,\ldots,H-1\}$. A patient of age t years who enters the state DOBC will stay in state DOBC with probability one, i.e., $T^{t}_{\text{DOBC, DOBC}}(a,\text{N}) = 1$ for all $a \in \mathcal{A}$ and $t \in \{0,1,\ldots,H-1\}$. Similar to a patient who is in the state DOBC, a patient who is in state PD will stay in state PD with probability one, i.e., $T^{t}_{\text{PD, PD}}(a,\text{N}) = 1$ for all $a \in \mathcal{A}$ and $t \in \{0,1,\ldots,H-1\}$.

When the chosen action is to wait until the next decision epoch, i.e., a^{t} =NS, only a patient in the state C can receive an abnormal observation result as shown in Table 3.2a. The patient makes a transition to the state PD with probability one after receiving the abnormal observation. Table 3.3 summarizes the transition probability matrix given the observation result is normal at a decision epoch t.

Table 3.3: The one-step transition matrix $T^{t}(a,\text{N})$ for a woman at age t given the normal observation

	NBC	P	C	DOBC	PD
NBC	$T^{t}_{\text{NBC,NBC}}$	$T^{t}_{\text{NBC,P}}$	0	T^{t}_{DOBC}	0
P	0	$T^{t}_{\text{P,P}}$	$T^{t}_{\text{P,C}}$	T^{t}_{DOBC}	0
C	0	0	0	0	1
DOBC	0	0	0	1	0
PD	0	0	0	0	1

When the observation result is abnormal, the transition probability matrix is more complicated. Recall from Figure 3.2, if the chosen action at decision epoch t is to wait and not to screen (NS), then the next possible event is to decide whether to screen at the next decision epoch $t+1$, which also implies that the transition matrix $T^{t}(\text{NS, Ab})$ is equivalent to $T^{t}(\text{NS, N})$. For a patient who is in the state NBC or P, because the observation result has probability equal zero, i.e., $O^{t}_{i,\text{Ab}}(\text{NS}) = 0$ for $i \in \{\text{NBC, P}\}$ and $t \in \mathbb{T} \setminus \{H\}$, the two rows in

the one-step transition matrix, $T_{i,j}^{\ t}(\text{NS, Ab})$ for $i \in \{\text{NBC, P}\}$ and $j \in \Omega$, which represent the process is currently in either state NBC or P, do not affect the process. Hence, for simplicity, the two rows are set to be equal to the two rows in the one-step transition matrix given the normal observation ℓ =N and the no screening action a =NS, i.e., $T_{i,j}^{\ t}(\text{NS, Ab}) = T_{i,j}^{\ t}(\text{NS, N})$ for $i \in \{\text{NBC, P}\}$ and $j \in \Omega$. On the other hand, for a patient who is in the state C, the observation result must be abnormal as shown in Eq. (3.3). Hence, the patient will move from state C to state PD with probability one.

As discussed in Section 3.1, improvements in breast cancer treatment and early detection have contributed greatly to the decline in breast cancer mortality. The main purpose of a mammographic test is to detect breast cancer early so that the patient who is in the state P can seek treatment options given the observation result is abnormal while a cure is more likely and the treatment is more effective. In other words, the action, to take a screening test (S), aims to enable the transition from state P to PD. Hence, the first four elements in the second row $T_{\text{P},j}^{\ t}(\text{S, Ab})$, $j \in \{\text{NBC, P, C, DOBC}\}$ of the transition matrix $T^{\ t}(\text{S, Ab})$ are equal to 0, and the last element $T_{\text{P, PD}}^{\ t}(\text{S, Ab})$ is equal to 1 because when the patient is in the preclinical state and screening yields an abnormal result, the follow-up diagnostic test will certainly detect the presence of cancer so that the patient moves to the postdiagnostic state with probability 1. Every other row holds the same in $T^{\ t}(\text{S, Ab})$ as it is in $T^{\ t}(\text{a, N})$. Table 3.4 summarizes the transition probability matrix given the observation result is abnormal.

When the observation result is one of the two observations other than the result of the mammogram, either Death or PD, which are added to capture that the process is in either one of the absorbing states, for simplicity, the one-step transition matrix is set to be equal to the one-step transition matrix when the action is to screen and the observation result is abnormal as shown in Table 3.4b. Although this definition of the one-step transition matrix $T^t(a^t, z^t)$ for

$z^t \in \{\text{Death, PD}\}$ may seem arbitrary, the associated observation probabilities

$$O^t_{i,\ell}(a^t) = 0 \text{ for } i \in \Omega', \ell \in \{\text{Death, PD}\}, a^t \in \mathcal{A}, \text{ and } t \in \mathbb{T} \setminus \{H\}$$

guarantee that all relevant computations involving $T^t(a^t, z^t)$ for $z^t \in \{\text{Death, PD}\}$ will yield the correct results. For the observation result that is associated with one of the absorbing states, the row which is associated with the absorbing state the transition matrix will lead the process to stay in the same absorbing state for the remaining decision epochs.

Table 3.4: The one-step transition matrix $T^{\,t}(a, \text{Ab})$ for a woman at age t given the abnormal observation

(a) *No Screening :* $a^{\,t}$ =NS

	NBC	P	C	DOBC	PD
NBC	$T^t_{\text{NBC,NBC}}$	$T^t_{\text{NBC,P}}$	0	T^t_{DOBC}	0
P	0	$T^t_{\text{P,P}}$	$T^t_{\text{P,C}}$	T^t_{DOBC}	0
C	0	0	0	0	1
DOBC	0	0	0	1	0
PD	0	0	0	0	1

(b) *Screening :* $a^{\,t}$ =S

	NBC	P	C	DOBC	PD
NBC	$T^t_{\text{NBC,NBC}}$	$T^t_{\text{NBC,P}}$	0	T^t_{DOBC}	0
P	0	0	0	0	1
C	0	0	0	0	1
DOBC	0	0	0	1	0
PD	0	0	0	0	1

Rewards

As discussed in Section 2.4, the immediate rewards depend on the state, the taken action, and the observation. In our breast cancer screening model, the reward we analyze is the lifetime breast cancer mortality probability associated with being detected in a particular breast cancer state at age t; and the purpose of the mammography is to detect breast cancer early so that the patient can seek treatment while breast cancer is most treatable so that the breast cancer mortality probability is reduced. Therefore in contract to the setup in Section 2.4 for a POMDP

model of general disease screening, our formulation of the POMDP model of breast cancer screening involves minimization rather than maximization of the value function at the beginning of the time horizon. Hence, if the observation is normal, then the immediate reward for each state is equal to zero because no breast cancer is detected given the normal observation result regardless of which action is chosen and there is no cost for a mammogram.

For $i \in \{\text{P, C}\}$, let $R^t_{i,\text{Ab}}$ denote the lifetime breast cancer mortality associated with being detected in the state P or C at a certain age t. For $i \in \{\text{P, C}\}$, the quantity $R^t_{i,\text{Ab}}$ is a lump-sum reward that includes all future rewards accumulated by the patient after she is detected with breast cancer at a certain stage (P or C). Hence, for the screening action, $a^t =$S, the immediate reward $q^t_{i,\text{Ab}}(a^t)$, is $R^t_{i,\text{Ab}}$ for $i \in \{\text{P, C}\}$ and $t \in \mathbb{T} \setminus \{H\}$. For the action not to screen, $a^t =$NS, the immediate rewards of being in states NBC or P with an abnormal observation result are equal to zero because the abnormal result will not happen given the action, not to screen, in the two states. However, the immediate result of being in state C with the abnormal observation result $R^t_{\text{C, Ab}}$ is the same as if the taken action is to screen and the observation result is abnormal.

When observation z^t is abnormal for both actions $a^t \in \mathcal{A}$, the mortality associated with being in the state, DOBC or PD, is 0. The zero mortality probability of being in the state DOBC is relatively straightforward. For a patient who is in state DOBC, she will incur neither any breast cancer mortality for the current decision epoch nor any lifetime breast cancer mortality from future decision epochs.

Once the patient moves into the state PD, the process will stay in the same state for the remaining decision epochs. Given the value function, Eq. (3.5b) below, the lifetime breast cancer mortality has been captured by $R^t_{i,\text{Ab}}$ for $i \in \{\text{P,C}\}$. Hence, the reward of being in the state PD is set to be zero.

For the immediate rewards associated with the observations, Death and PD, the reward has been captured by $R^t_{i,\text{Ab}}$ for $i \in \{\text{P, C}\}$ when the breast cancer is detected. Hence, the immediate

rewards are set to be zero given the observation equal to either Death or PD, $O^t_{i,\ell}(a^t)$ for $i \in \Omega$, $\ell \in \{\text{Death, PD}\}$, and $a \in \mathcal{A}$. Table 3.5 summarizes the immediate rewards in a matrix form.

Table 3.5: The immediate reward $q^t_{i,\ell}(a)$ for a woman at age t given the chosen action

(a) *Not to screen :* a^t =NS

$$x^t = \begin{array}{c} \\ \text{NBC} \\ \text{P} \\ \text{C} \\ \text{DOBC} \\ \text{PD} \end{array} \overbrace{\begin{array}{cccc} \text{N} & \text{Ab} & \text{Death} & \text{PD} \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & R^t_{\text{C, Ab}} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}}^{z^t=}$$

(b) *Screening :* a^t =S

$$x^t = \begin{array}{c} \\ \text{NBC} \\ \text{P} \\ \text{C} \\ \text{DOBC} \\ \text{PD} \end{array} \overbrace{\begin{array}{cccc} \text{N} & \text{Ab} & \text{Death} & \text{PD} \\ \hline 0 & 0 & 0 & 0 \\ 0 & R^t_{\text{P, Ab}} & 0 & 0 \\ 0 & R^t_{\text{C, Ab}} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}}^{z^t=}$$

For a woman of age H, the vector of terminal rewards, q^{H}, is given by

$$q^{H} = \begin{bmatrix} 0 \\ R_{\mathrm{P}}^{H} \\ R_{\mathrm{C}}^{H} \\ 0 \\ 0 \end{bmatrix}. \tag{3.4}$$

Eq. (3.4) states that, at the end of the time horizon H, if a patient is in one of the breast cancer states, P or C, then the value function should add some reward, i.e., the lifetime breast cancer mortality probability associated with being detected in a particular breast cancer state at age H. If the patient is in one of the absorbing states, DOBC or PD, then the value function will not add any lifetime breast cancer mortality probability. If a patient is in the state NBC when she reaches age 100, the probability of dying from other causes, is approximately 1, so that the probability of dying from breast cancer beyond age 100 is assumed to be 0 when a patient has no breast cancer at age 100.

Given the updating formulas, the observation matrices, and the transition matrices, we can calculate the sample path of a selected initial belief state and a policy. Figure 3.4 shows an example of a three-period sample path for a patient of age 70 with the beginning state $\pi^{0} = [0.90, 0.08, 0.02, 0, 0]$, and following the screening policy that has 2 screenings at the first and the last decision epochs and no screening at the second decision epoch, $\varphi = [\mathrm{S}, \mathrm{NS}, \mathrm{S}]$. The notation in Figure 3.4 is similar to the sample path example in Figure 2.4. A node represents a belief state in which the process will be at the beginning of a particular decision epoch. The number pair inside a node differentiates the nodes from each other. The first element in the pair represents the decision epoch and the second element in the pair represents the *i*th possible

outcome at the same decision epoch. Each directed branch (edge, arc) connects the current belief state (origin node) and the next possible belief state (destination node) given the taken action and the possible observation. The pair associated with each branch represents the taken action and the observed result. Each a node can have several branches originating at that node.

If the taken action is to screen, then as discussed earlier in the observation part, the patient will not take any action after receiving the normal observation result and will follow the natural progression until the next screening or until breast cancer symptoms develop as shown in nodes (1, 1), (2, 1), (2, 2), (2, 3), and (2, 4). On the other hand, if the observation result is abnormal, then the patient will act differently depending on if the result is a true positive or a false positive. If the abnormal result is a true positive, i.e., the process is in either state P or C, then the process will enter state PD. On the other hand, if the abnormal result is a false positive result, i.e., the process is in state NBC, then the process will transition either to state NBC, P, or D as if the belief of the patient's health were reset to be in state NBC with probability one. The belief state associated with node (1, 2) represents the probability after an abnormal observation at decision epoch $t = 0$ and a transition into different states at beginning of decision epoch $t = 1$.

The abnormal observation result given the chosen action is not to screen (the arrows labelled (NS, Ab)) only occurs if the belief state at beginning of the current decision epoch t has a positive probability of being in state C. Hence, the belief state at the next decision epoch $t+1$ can only be in state PD with probability one, $[0,0,0,0,1]$. Note that the nodes with belief state equal to DOBC or PD with probability one will stay in the same belief state, because state DOBC and PD are fully observable and absorbing. Hence, to simplify the graph, no succeeding node is presented in the graph.

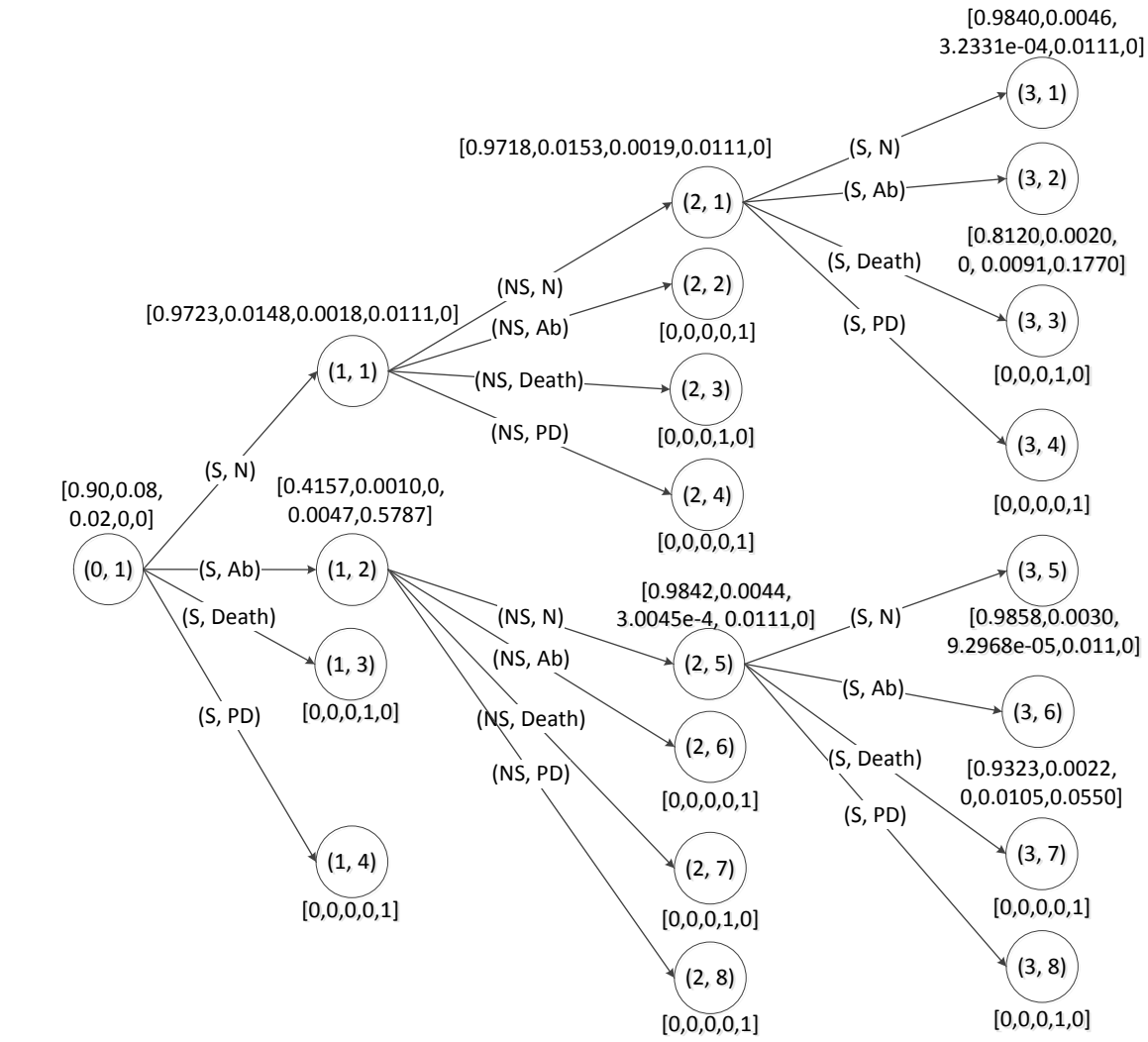


Figure 3.4: A three-period sample path example for a patient at age 70 with the beginning state, $[0.90, 0.08, 0.02, 0, 0]$, and the screening policy that has 2 screenings at the first and the last decision epochs and no screening at the second decision epoch, $\varphi = [\text{S, NS, S}]$.

The Value Function and The Policy

For a given policy $\varphi : (\pi^{\,t}, t) \in \Pi \times \mathbb{T} \mapsto a = \varphi(\pi^{\,t}, t) \in \mathcal{A}$, we can write the value function at decision epoch $t \in \mathbb{T}$ as follows:

$$V^{\,H}(\pi) = \sum_{i \in \Omega} \pi_i q_i^{\,H}, \tag{3.5a}$$

$$V^{\,t}(\pi^{\,t}, \varphi) = \sum_{i \in \Omega} \pi_i{}^{t} \sum_{\ell \in \mathcal{Z}} O_{i,\ell}^{t}\left(\varphi(\pi^{\,t}, t)\right) \left\{ q_{i,\ell}^{t}(\varphi(\pi^{\,t}, t)) + \sum_{j \in \Omega} T_{i,j}^{t}\left(\varphi(\pi^{\,t}, t), \ell\right) \mathbb{V}_j^{\,t+1} \right\}$$
$$\text{for } t \in \mathbb{T} \setminus \{H\} \tag{3.5b}$$

where for simplicity we write

$$\mathbb{V}_j^{\,t+1} = V_j^{\,t+1}\left(\pi_j^{\,t+1}(\varphi(\pi^{\,t}, t), \ell), \varphi(\pi^{\,t+1}, t+1)\right) \text{ for } t \in \mathbb{T} \setminus \{H\}, \text{ and } j \in \Omega, \tag{3.5c}$$

where the auxiliary functions $\left\{ V_j^t(\pi_j^t, a^t) : j \in \Omega, t \in \mathbb{T} \setminus \{H\}, a^t \in \mathcal{A} \right\}$ are defined recursively as in Eq. (2.31) in Section 2.4. Eq. (3.5a) is the value function for the end of the time horizon, i.e., for a woman at age 100, while Eq. (3.5b) and Eq. (3.5c) are for all the other decision epochs.

Note that mammography seeks to reduce breast cancer mortality, and we use breast cancer mortality probabilities as the reward. Hence, the optimal policy is defined as the following,

$$\widetilde{\varphi}\left(\pi^{\,t}, t\right) = \underset{a \in \mathcal{A}}{\arg\min} \left\{ \sum_{i \in \Omega} \pi_i{}^{t} \sum_{\ell \in \mathcal{Z}} O_{i,\ell}^{t}\left(a^t\right) \left[q_{i,\ell}^{t}(a^t) + \sum_{j \in \Omega} T_{i,j}^{t}\left(a^t, \ell\right) \mathbb{V}_j^{\,t+1} \right] \right\} \text{ for } t \in \mathbb{T} \setminus \{H\}. \tag{3.6}$$

Notice that Eq. (3.6) exactly parallels Eq. (2.32) of Section 2.4, except that the former involves the arg min operator whereas the latter involves the arg max operator.

3.4 Inverse POMDP model for Breast Cancer Screening

In this section, we discuss the formulation of an inverse POMDP model for the breast cancer screening problem. To solve the inverse breast cancer POMDP model, we use the general inverse algorithm idea from Chapter 2 with some constraints that are specifically defined for this breast cancer screening application.

Similar to the general inverse POMDP problem formulation in Section 2.5, we will have a nonlinear programming problem for a chosen policy which has the same form as Eq. (2.42). Moreover, there are two additional conditions (Eq. (3.7e) and Eq. (3.7f) below) that need to be considered for the breast cancer screening problem. First, the value function result for each policy should be bounded between zero and one as shown in Eq. (3.7e). This follows directly because the reward is the lifetime breast cancer mortality probability: one can have neither a negative mortality probability nor a mortality probability which is greater than one.

The second condition is that each of the unknowns, $R^{t}_{\mathrm{P,\,Ab}}$ and $R^{t}_{\mathrm{C,\,Ab}}$, should be between zero and one as shown in Eq. (3.7f). Similar to the constraint Eq. (3.7e), the reason is one can have neither a negative mortality probability at a certain decision epoch t in one of the cancer states, state P or C, nor a mortality probability which is greater than one. Let $\mathbb{Q}$ denote all the unknowns, $R^{t}_{i,\mathrm{Ab}}$ for $i \in \{\mathrm{P,\,C}\}$, which form a subset of the entire reward functions $\mathcal{R}$.

Given the two conditions above and a time-dependent policy $\widetilde{\varphi}$ that is assumed to be optimal, we can form the nonlinear programming problem as following,

$$\max_{\substack{\pi^0 \in \Pi \\ \mathbb{Q}}} \sum_{\widehat{\varphi} \in \Phi \setminus \{\widetilde{\varphi}\}} \left\{ \widehat{V}^0\left(\pi^0, \widehat{\varphi}\right) - \widetilde{V}^0\left(\pi^0, \widetilde{\varphi}\right) \right\}, \tag{3.7a}$$

$$s.t. \quad \widehat{V}^0\left(\pi^0, \widehat{\varphi}\right) - \widetilde{V}^0\left(\pi^0, \widetilde{\varphi}\right) \geqslant 0, \text{ for all } \pi^0 \in \Pi \text{ and } \widehat{\varphi} \in \Phi \setminus \{\widetilde{\varphi}\}, \tag{3.7b}$$

$$\sum_{i \in \Omega} \pi^0_i = 1, \text{ for all } \pi^0 \in \Pi \tag{3.7c}$$

$$0 \leqslant \pi_i^0 \leqslant 1, \quad \text{for all } i \in \Omega, \text{ and } \pi^0 \in \Pi \tag{3.7d}$$

$$0 \leqslant V^0\left(\pi^0, \varphi\right) \leqslant 1, \quad \text{for all } \varphi \in \Phi, \text{ and } \pi^0 \in \Pi \tag{3.7e}$$

$$0 \leqslant R_{i,\text{Ab}}^t \leqslant 1, \quad \text{for } i \in \{\text{P, C}\} \text{ and } t \in \mathbb{T} \setminus \{H\}. \tag{3.7f}$$

Recall from Section 2.5, $\widetilde{V}^0(\pi^0, \widetilde{\varphi})$ represents the value function given the designated policy $\widetilde{\varphi}$, and $\widehat{V}^0(\pi^0, \widehat{\varphi})$ represents the value function given an nonoptimal policy $\widehat{\varphi} \in \Phi \setminus \{\widetilde{\varphi}\}$. Note that $V^0\left(\pi^0, \varphi\right)$ in Eq. (3.7e) represents the value function result of all policies, including $\widetilde{V}^0\left(\pi^0, \widetilde{\varphi}\right)$ and $\widehat{V}^0\left(\pi^0, \widehat{\varphi}\right)$.

Also, given the definition of the optimal policy Eq. (3.6), the value function for the optimal policy is smaller than for all the other policies. Hence, instead of taking

$$\sum_{\widehat{\varphi} \in \Phi \setminus \{\widetilde{\varphi}\}} \left\{ \widetilde{V}^0\left(\pi^0, \widetilde{\varphi}\right) - \widehat{V}^0\left(\pi^0, \widehat{\varphi}\right) \right\},$$

as shown in Eq. (2.42a), we take

$$\sum_{\widehat{\varphi} \in \Phi \setminus \{\widetilde{\varphi}\}} \left\{ \widehat{V}^0\left(\pi^0, \widehat{\varphi}\right) - \widetilde{V}^0\left(\pi^0, \widetilde{\varphi}\right) \right\},$$

so that the objective function is greater than zero.

As mentioned in Section 2.5, we discretize the entire beginning state space into a finite fixed grid $\mathscr{G} \subset \Pi$ so that the problem, Eq. (3.7), is converted from a nonlinear problem to several linear problems given the state space constraints, Eq. (3.7c) and Eq. (3.7d). Therefore, the linear problem given the initial belief state $\pi^0 = \pi_x \in \mathscr{G}$ take the following form:

$$\max_{\mathbb{Q}} \sum_{\widehat{\varphi} \in \Phi \setminus \{\widetilde{\varphi}\}} \left\{ \widehat{V}^0\left(\pi_x, \widehat{\varphi}\right) - \widetilde{V}^0\left(\pi_x, \widetilde{\varphi}\right) \right\}, \tag{3.8a}$$

$$s.t. \quad \widehat{V}^0(\pi_x, \widehat{\varphi}) - \widetilde{V}^0(\pi_x, \widetilde{\varphi}) \geqslant 0, \text{ for } \widehat{\varphi} \in \Phi \setminus \widetilde{\varphi}, \tag{3.8b}$$

$$0 \leqslant V^0(\pi_x, \varphi) \leqslant 1, \quad \text{for all } \varphi \in \Phi, \tag{3.8c}$$

$$0 \leqslant R^t_{i,\text{Ab}} \leqslant 1, \quad \text{for } i \in \{\text{P, C}\} \text{ and } t \in \mathbb{T} \setminus \{H\}. \tag{3.8d}$$

Hence, we solve a linear problem for every initial belief state in the finite fixed grid, $\pi_x \in \mathscr{G}$ and combine the results for each initial belief state $\pi_x \in \mathscr{G}$ by taking the average imputed rewards over the finite fixed grid as shown in Eq. (2.44).

To avoid the operational infeasibility problem as discussed in Section 2.5, the age group concept is also applied to this breast cancer screening policy inverse problem. We select five years as the length of each age group, i.e., ten decision epochs for an age group, $\mathcal{L} = 10$, so that the age groups of the inverse problem matches the available data of the transition matrix and the observation matrix. In this way, we can eliminate the variability of the transition matrix and the observation matrix while testing the model. According to Eq. (2.45), we have 15 different age groups, i.e., 15 different sets of unknowns, $R^{\alpha(t)}_{i,\text{Ab}}$, $\alpha(t) \in \{1, 2, \ldots, 15\}$, $i \in \{\text{P, C}\}$, to solve. Combined with the robust method shown in Eq. (2.46), the inverse problem is as the following,

$$\max_{\mathbb{Q}} \sum_{\widehat{\varphi} \in \Phi \setminus \{\widetilde{\varphi}\}} \left\{ \widehat{V}^0(\pi_x, \widehat{\varphi}) - \widetilde{V}^0(\pi_x, \widetilde{\varphi}) \right\}, \tag{3.9a}$$

$$s.t. \quad \widehat{V}^0(\pi_x, \widehat{\varphi}) - \widetilde{V}^0(\pi_x, \widetilde{\varphi}) \geqslant 0, \text{ for } \widehat{\varphi} \in \Phi \setminus \{\widetilde{\varphi}\}, \tag{3.9b}$$

$$0 \leqslant V^0(\pi_x, \varphi) \leqslant 1, \quad \text{for all } \varphi \in \Phi, \tag{3.9c}$$

$$0 \leqslant R^t_{i,\text{Ab}} \leqslant 1, \quad \text{for } i \in \{\text{P, C}\} \text{ and } t \in \mathbb{T} \setminus \{H\}, \tag{3.9d}$$

$$R^\tau_{i,\text{Ab}} - \varepsilon \leqslant R^\beta_{i,\text{Ab}} \leqslant R^\tau_{i,\text{Ab}} + \varepsilon,$$

$$\text{for } \beta \in \{\tau + 1, \tau + 2, \ldots, \tau + \mathcal{L} - 1\} \text{ and } i \in \{\text{P, C}\}, \tag{3.9e}$$

$$\tau = \mathcal{L}\left(\alpha(t) - 1\right) + 1, \quad \text{for } t \in \{1, 2, \ldots, H-1\}. \tag{3.9f}$$

3.5 Conclusion

In this chapter, we present a forward POMDP for finding an optimal time-dependent breast cancer screening policy. A Markov chain serves as the core process to describe the natural progression of breast cancer before the patient is detected with the disease. Also, details of the forward POMDP formulation are discussed in Section 3.3.2.

Given the forward POMDP for finding an optimal time-dependent breast cancer screening policy, we present the formulation of the inverse problem for each decision epoch. In addition to the constraints presented in Section 2.5, we add several constraints designed for this breast cancer screening POMDP model. First, the rewards of the forward POMPD are the lifetime breast cancer mortality probabilities upon detecting the breast cancer in different stages, pre-clinical and clinical. Therefore, the first constraint is that the rewards are constrained to fall between zero and one. Also, given the definition of the rewards, the result of the value function should also between zero and one which forms the second constraint.

Chapter 4

The Imputed Rewards of Breast Cancer Screening Policies

In Section 4.1, we discuss the design of the computational experiment used to evaluate our procedure for imputing the rewards (mortality probabilities) associated with a given breast cancer screening policy. Then, the method to evaluate the breast cancer screening policies is presented in Section 4.2. In Section 4.3, we explain in detail the data used in our model. The numerical results are presented in Section 4.4. Finally, in Section 4.5, we summarize the most important conclusions for this chapter.

4.1 Experiment Design

As mentioned in Section 2.6, the computational effort required to solve the inverse problem is not attractive. Also, given the data limitation that the transition probability matrix and the observation probability matrix are actually estimated for several different age groups, the forward POMDP problem is actually formed by several small problems that are linked together. There-

fore, to solve the inverse POMDP for the problem of finding the optimal time-dependent policy for breast cancer screening, we divide the whole time horizon into several small segments with the length equal to five years. In this way, the length of each segment matches the following:

- the age group length, $\mathcal{L} = 10$ decision epochs with a period of length six months between successive decision epoch as mentioned in Section 3.4; and
- the available transition matrices and the available observation matrices.

We select five different age groups to form small inverse problems for solving the breast cancer screening policy problem described in Section 3.4. The five age groups are ages 30–34, 40–44, 50–54, 75–79, and 85–89. The two age groups, ages 40–44 and 50–54, are the beginning age groups of the breast cancer screening policies recommended by the ACS and the USPSTF, respectively; and the last two age groups are selected to represent the age groups that are not recommended for screening tests.

Within each age group problem, Φ consists of 1024 policies ($\kappa^{\mathcal{L}} = 2^{10} = 1024$). The overall experiment consists of the following steps:

Step 1 Select the grid $\mathscr{G}$ of initial belief states $\pi_x \in \mathscr{G} \subset \Pi$ as described below;

Step 2 Solve the LP problem Eq. (3.9) for each policy $\varphi \in \Phi$ with each initial belief state $\pi_x \in \mathscr{G}$ by assuming the chosen policy φ is the optimal policy;

Step 3 Evaluate each policy $\varphi \in \Phi$ using the following performance measures: (a) $M^0(\varphi; \pi_x)$, the expected number of mammograms over the entire time horizon $\{0, 1, \ldots, H\}$ for the policy φ given $\pi_x \in \Pi$ at decision epoch 0; and (b) $\widetilde{V}^0(\pi_x, \widetilde{\varphi})$, the value function at decision epoch 0, which is evaluated with the imputed rewards, $R_{i,\mathrm{Ab}}^{\alpha(t)}, i \in \{\mathrm{P}, \mathrm{C}\}$ associated with the policy φ.

Selecting the initial belief state plays an important role in the inverse algorithm because the sample path of each policy relies on the initial belief state. Let $\ddot{\pi}_{\text{NBC}}$ denote the lower bound on π^t_{NBC} that the breast cancer natural process can enter during a patient's lifetime,

$$\ddot{\pi}_{\text{NBC}} = \min\left\{\pi^t_{\text{NBC}} : t \in \mathbb{T}\right\}. \tag{4.1}$$

To select the initial belief states, the estimate of a woman's risk of developing invasive breast cancer over her lifetime, *risk*, is found from the Breast Cancer Risk Assessment Tool from National Cancer Institute [33]. Given the estimate, $1 - risk$ is assumed to be $\ddot{\pi}_{\text{NBC}}$, the minimal probability of being in the state NBC at the beginning of an age group. From the assessment tool, we learned that *risk* can be as high as 81.5%, which means that $\ddot{\pi}_{\text{NBC}}$ can be as low as $1 - 81.5\% = 18.5\%$. Hence, we select 15% as the lowest probability of being in the state NBC, $\ddot{\pi}_{\text{NBC}} = 1 - risk = 15\%$.

Given the assumption that the breast cancer does not regress without any medical intervention, the only feasible initial belief state with $\pi^0_{\text{P}} = 0$ is the belief state $\pi^0 = [1,0,0,0,0]$. Hence, an initial belief state with the form, $\pi^0 = [\pi_{\text{NBC}}, \pi_{\text{P}} = 0, \pi_{\text{C}}, \pi_{\text{DOBC}}, \pi_{\text{PD}}] \in \Pi$, is not included in the grid $\mathscr{G}$. For example, an initial belief state $\pi^0 = [0.80,0,0.20,0,0]$ is not considered as an initial belief state to test the inverse algorithm. Also, the initial belief states with a positive probability in either one of the absorbing states are not included in the initial belief state grid $\mathscr{G}$ because the focus of this research is on women who are still alive and have not already been detected with breast cancer. Hence, Table 4.1 shows the initial belief states in the grid $\pi^0 \in \mathscr{G}$.

Table 4.1: The list of the initial belief states in the grid $\mathscr{G} = \{\pi^0(j) : j = 1, \ldots, 37\}$.

	Components of $\pi^0(j)$				
j	π_{NBC}	π_{P}	π_{C}	π_{DOBC}	π_{PD}
1	0.2	0.1	0.7	0	0
2	0.2	0.2	0.6	0	0
3	0.2	0.3	0.5	0	0
4	0.2	0.4	0.4	0	0
5	0.2	0.5	0.3	0	0
6	0.2	0.6	0.2	0	0
7	0.2	0.7	0.1	0	0
8	0.2	0.8	0	0	0
9	0.3	0.1	0.6	0	0
10	0.3	0.2	0.5	0	0
11	0.3	0.3	0.4	0	0
12	0.3	0.4	0.3	0	0
13	0.3	0.5	0.2	0	0
14	0.3	0.6	0.1	0	0
15	0.3	0.7	0	0	0
16	0.4	0.1	0.5	0	0
17	0.4	0.2	0.4	0	0
18	0.4	0.3	0.3	0	0
19	0.4	0.4	0.2	0	0
20	0.4	0.5	0.1	0	0
21	0.4	0.6	0	0	0
22	0.5	0.1	0.4	0	0
23	0.5	0.2	0.3	0	0
24	0.5	0.3	0.2	0	0
25	0.5	0.4	0.1	0	0
26	0.5	0.5	0	0	0
27	0.6	0.1	0.3	0	0

Table 4.1: (Continued)

	Components of $\pi^0(j)$				
j	π_{NBC}	π_{P}	π_{C}	π_{DOBC}	π_{PD}
28	0.6	0.2	0.2	0	0
29	0.6	0.3	0.1	0	0
30	0.6	0.4	0	0	0
31	0.7	0.1	0.2	0	0
32	0.7	0.2	0.1	0	0
33	0.7	0.3	0	0	0
34	0.8	0.1	0.1	0	0
35	0.8	0.2	0	0	0
36	0.9	0.1	0	0	0
37	1.0	0	0	0	0

Not only the total expected lifetime breast cancer mortality probability $V^0(\pi^0, \varphi)$ for $\pi^0 \in \Pi$ but also the effort (cost) of executing a policy should be taken into account when evaluating a screening policy $\varphi \in \Phi$ so that an "optimal" screening policy can most efficiently reduce the lifetime breast cancer mortality probability — i.e., for two policies that yield equivalent reductions in total expected lifetime breast cancer mortality probability, the policy with the smaller expected number of mammograms is the more efficient policy. The expected number of mammograms is introduced to evaluate the effort of executing a policy in the next section.

4.2 Evaluation Method

For a given policy $\varphi \in \Phi$, belief state $\pi^t \in \Pi$, and the action $a^t \in \mathcal{A}$ at decision epoch t, $M^t(\varphi; \pi^t)$ is defined to be the conditional expected value of the number of mammograms that a woman will receive by adhering to policy φ from decision epoch t to decision epoch $H-1$, given that she is in belief state π^t at decision epoch t. In a nutshell, the calculation of the expected number of mammograms is similar as the value function with the immediate rewards as shown in Table 4.2. Let $m^t_{i,\ell}(a^t)$ denote the immediate number of mammograms given the core process is in state $i \in \Omega$, the action $a^t \in \mathcal{A}$ is taken, and the observation $\ell \in \mathcal{Z}$ is received at decision epoch $t \in \mathbb{T} \setminus \{H\}$. When the action is not to screen so that $a^t =$NS, the immediate number of mammograms is set to be zero, $m^t_{i,\ell}(a^t = \text{NS}) = 0$ for $i \in \Omega$ and $\ell \in \mathcal{Z}$. On the other hand, when the action is to screen so that $a^t =$S, only a patient who is in a transient state will take the screening test, $m^t_{i,\ell}(a^t = \text{NS}) = 1$ for $i \in \Omega'$ and $\ell \in \{\text{N}, \text{Ab}\}$. Note that the two observation results, Death and PD, are added to represent the situation in which the core process enters one of the absorbing states and will stay in that absorbing state for the remaining decision epochs. Hence, the immediate number of mammograms associated with these two observations are set to be zero even if a patient is in one of the transient states,

$m^t_{i,\ell}(a^t = \mathrm{S}) = 0$ for $i \in \Omega'$ and $\ell \in \{\text{Death, PD}\}$. For a patient who is in state DOBC or PD, the immediate number of mammograms are set to be zero $m^t_{i,\ell}(a^t = \mathrm{S}) = 0$ for $i \in \Omega \setminus \Omega'$ and $\ell \in \mathcal{Z}$.

Table 4.2: The immediate number of mammograms, $m^t_{i,\ell}(a^t)$ for $i \in \Omega$, $\ell \in \mathcal{Z}$, $a \in \mathcal{A}$, and $t \in \mathbb{T} \setminus \{H\}$

(a) *Not to screen :* a^t =NS

z^t=

x^t	N	Ab	Death	PD
NBC	0	0	0	0
P	0	0	0	0
C	0	0	0	0
DOBC	0	0	0	0
PD	0	0	0	0

(b) *Screening :* a^t =S

z^t=

x^t	N	Ab	Death	PD
NBC	1	1	0	0
P	1	1	0	0
C	1	1	0	0
DOBC	0	0	0	0
PD	0	0	0	0

For a given policy $\varphi \in \Phi$, the performance measure $M^t(\varphi;\pi^t)$ is calculated recursively for $t = H, H-1, \ldots, 1, 0$, as follows:

$$M^H(\varphi;\pi^H) = 0, \text{ for all } \pi^H \in \Pi, \tag{4.2a}$$

$$\begin{aligned} M^t(\varphi;\pi^t) = \sum_{i\in\Omega} \pi^t_i \Bigg\{ & \sum_{\ell\in\mathcal{Z}} O^t_{i,\ell}(\varphi(\pi^t,t)) m^t_{i,\ell}(\varphi(\pi^t,t)) \\ & + \sum_{\ell\in\mathcal{Z}} \sum_{j\in\Omega} O^t_{i,\ell}\left(\varphi(\pi^t,t)\right) T^t_{i,j}\left(\varphi(\pi^t,t),\ell\right) M^{t+1}_j\left(\varphi;\pi^{t+1}_j\left(\varphi(\pi^t,t),\ell\right)\right) \Bigg\} \\ & \text{for } \pi^t \in \Pi, \varphi \in \Phi, \text{ and } t \in \mathbb{T} \setminus \{H\}, \end{aligned} \tag{4.2b}$$

where the auxiliary functions $\{M^t_j(\varphi;\pi^t) : i \in \Omega, \text{ and } t = H, H-1, \ldots, 1, 0\}$ are defined recur-

sively as follows:

$$M_i^H(\varphi;\pi_i^H) = 0 \text{ for } i \in \Omega, \varphi \in \Phi, \text{ and } \pi^H \in \Pi, \tag{4.2c}$$

$$M_i^t(\varphi;\pi_i^t) = {\pi_i}^t \left\{ \sum_{\ell\in\mathcal{Z}} O_{i,\ell}^t(\varphi(\pi^t,t)) {m_{i,\ell}}^t(\varphi(\pi^t,t)) + \sum_{\ell\in\mathcal{Z}} \sum_{j\in\Omega} O_{i,\ell}^t(\varphi(\pi^t,t)) {T_{i,j}}^t(\varphi(\pi^t,t),\ell) M_j^{t+1}\left(\varphi;\pi_j^{t+1}\left(\varphi(\pi^t,t),\ell\right)\right) \right\}$$

$$\text{for } \pi^t \in \Pi, i \in \Omega, \varphi \in \Phi, \text{ and } t \in \mathbb{T}\setminus\{H\}, \tag{4.2d}$$

and where as usual $\pi^{t+1}(\varphi(\pi^t,t),\ell)$ is given by the Bayesian updating formula Eq. (2.30) for $t \in \mathbb{T}\setminus\{H\}$, $a^t \in \mathcal{A}$, and $\ell \in \mathcal{Z}$.

Using the three-period sample path example shown in Figure 4.1, a patient at age 70 with the belief state $[0.90, 0.08, 0.02, 0, 0]$ follows the screening policy that has 2 screenings at the first and the last decision epochs and no screening at the second decision epoch. The numbers outside of the nodes represent the expected number of mammograms at the particular decision epoch. For the nodes labelled (3, ℓ) for $1 \leqslant \ell \leqslant 8$, the expected number of mammograms is equal to zero as shown in Eq. (4.2a). For all other nodes, the calculation follows Eq. (4.2b) and Eq. (4.2d). Therefore, at the beginning of the three periods, this patient is expected to receive 1.7689 mammograms.

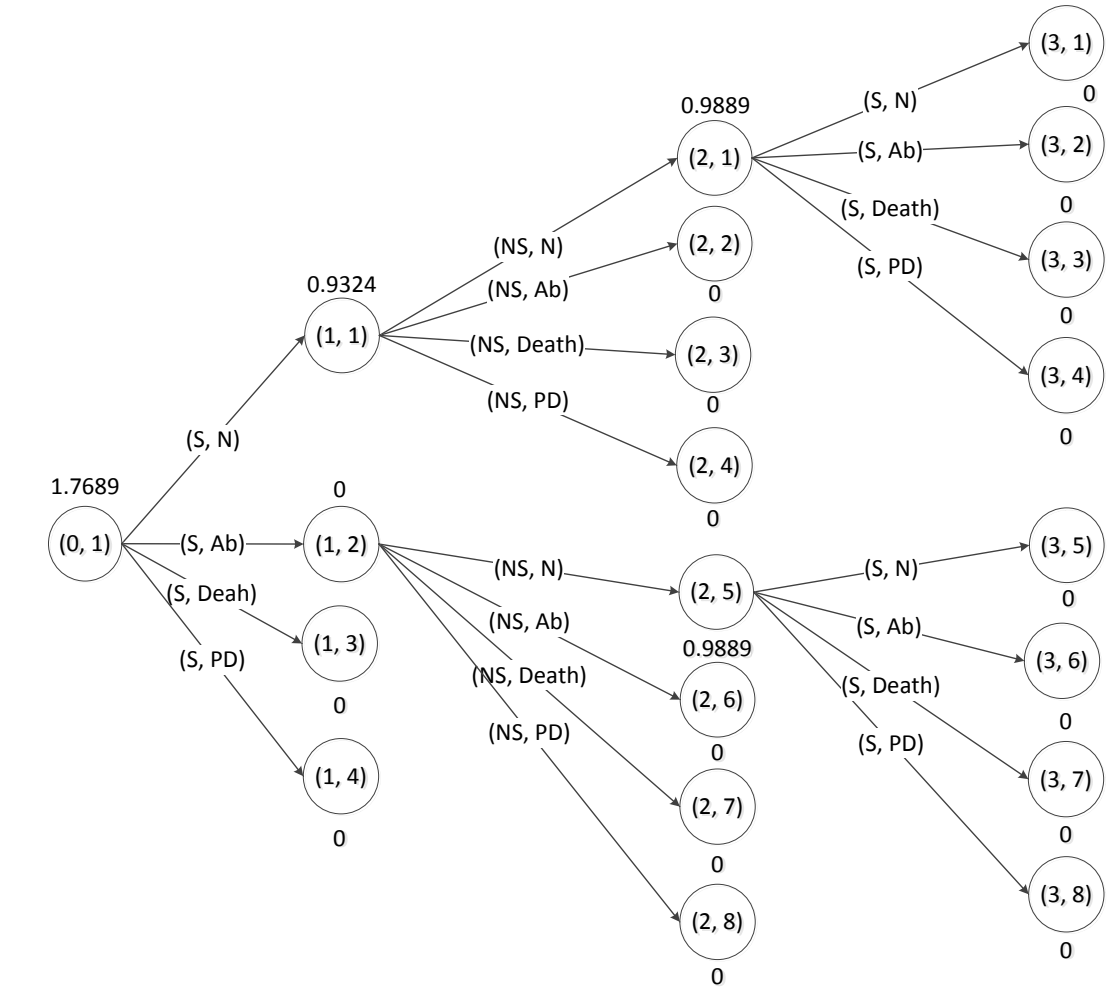


Figure 4.1: A three-period expected number of mammograms example for a patient at age 70 with the beginning state, $[0.90, 0.08, 0.02, 0, 0]$, and the screening policy that has 2 screenings at the first and the last decision epochs and no screening at the second decision epoch, $\varphi =$ [S, NS, S].

4.3 Data Description

In this section, we provide the data sources for the numerical experiments. Also, we present the methods of estimating all parameters from the available data.

Table 4.3 summarizes the data sources for our model parameters. A woman's health progression changes with age, i.e., the speed of breast cancer development is different for women at different ages. Ideally, the transition matrix $T^{\,t}$ should be a function of the decision epoch t, which corresponds to the five-year age group with beginning age $(\mathcal{L}/2) \times (\alpha(t) - 1) + 25$ expressed in years, where $\alpha(t) = \lceil t/\mathcal{L} \rceil + 1$ for $t \in \{0, 1, \ldots, H-1\}$ as defined in Eq. (2.45). Due to limitations in the available data, we cluster every five years as an age group for women between ages 25 and 85, and the last age group contains women from age 85 to age 100. Thus, thirteen age groups are included in the model, ages 25 to 29, ages 30 to 34, ages 35 to 39, ages 40 to 44, ages 45 to 49, ages 50 to 54, ages 55 to 59, ages 60 to 64, ages 65 to 69, ages 70 to 74, ages 75 to 79, ages 80 to 84, and ages 85 to 100. We estimate the model parameters according to the age groups.

The mean sojourn time (MST), measured in years, is used to calculate the parameter $T_{\mathrm{P,\,C}}^{\,t}$. Recall that sojourn time is the time interval in Figure 3.1 with the label "Detectable preclinical phase of breast cancer." As discussed in Section 3.3, MST refers to the time the breast cancer stays in the preclinical phase, after the biologic onset of breast cancer and before symptoms develop. We use the following formula from Maillart et al. [30] to evaluate the parameter, $T_{\mathrm{P,\,C}}^{\,t}$,

$$T_{\mathrm{P,\,C}}^{\,t} = \frac{1}{2\mathrm{MST}^{\,t}}. \tag{4.3}$$

The multiplication of one half by the inverse of the MST in Eq. (4.3) converts the time unit from one year to six months. We use the estimates of Tabár et al. [46] for the MSTs of the

Table 4.3: Data source for model parameter estimation

Model parameters for $t \in \mathbb{T}$		Data source
Aggression	$T^{t}_{\mathrm{P,\,C}}$	[46]
		[51]
Incidence	$T^{t}_{\mathrm{NBC,\,P}}$	[37]
Comorbidity	$T^{t}_{\mathrm{NBC,\,DOBC}}$	[18]
	$T^{t}_{\mathrm{P,\,DOBC}}$	[18]
Specificity	$O^{t}_{\mathrm{NBC,\,Ab}}$	[52]
		[15]
Sensitivity	$O^{t}_{\mathrm{P,\,Ab}}$	[52]
		[15]
Lifetime Mortality by Age and Stage	$R_{i,\mathrm{Ab}}^{(\mathcal{L}/2)\times(\alpha(t)-1)+25}, i \in \{\mathrm{P,\,C}\}$	[55], [12]

following age groups: ages 40 to 49, ages 60 to 69, and ages 70 to 79; and we use the estimates of Wu et al. [51] for the MSTs for ages 50 to 54 and ages 55 to 59. For the younger age groups and the older age groups, we assume the mean sojourn time remains the same as for the closest available age group. In other words, we use the mean sojourn time of the age group 40 to 49 to fill in the mean sojourn time of the age groups 25 to 29, 30 to 34, and 35 to 39; and we use the mean sojourn time of the age group 70 to 79 to fill in the mean sojourn time of the age groups 80 to 85 and 85 to 100.

The incidence rate $T^{t}_{\mathrm{NBC,\,P}}$ is acquired from the online statistics database tool on the Surveillance Epidemiology and End Results (SEER) website [37]. We use only the in situ breast cancer incidence rate to estimate the incidence parameter for each age group, which is the best fit based on our state space definition from the available data. In situ breast cancer is only detectable by mammogram and is, by definition, preclinical. However, some invasive breast cancers are also asymptomatic, so this assumption may underestimate incidence.

We use the data from the Centers for Disease Control (CDC) WONDER online database

[18] to estimate the parameters $T^{t}_{\text{NBC, DOBC}}$ and $T^{t}_{\text{P, DOBC}}$ for all age groups. From the online database, we found the number of female deaths from all causes, D^{t}, and the number of deaths from female breast cancer, DBC^{t}, for each $t \in \mathbb{T}$ and hence for each age group. The size of the female population in each age group, ξ^{t}, is drawn from the same database. We assume that two parameters, $T^{t}_{\text{NBC, DOBC}}$ and $T^{t}_{\text{P, DOBC}}$, are equal to each other because, for a patient who is in the state P, she will not die from breast cancer without progressing to state C in a one-step transition, so that the probability of dying from other causes while she is in the state P is equal to the probability of dying from other causes when she is in the state NBC. The following equation is used to calculate the one-step transition probabilities of moving from state NBC or P to death from causes other than breast cancer (state DOBC),

$$T^{t}_{\text{NBC, DOBC}} = T^{t}_{\text{P, DOBC}} = \frac{\text{D}^{t} - \text{DBC}^{t}}{\xi^{t}} \quad \text{for } t \in \mathbb{T}. \tag{4.4}$$

The two remaining parameters, $T^{t}_{\text{NBC, NBC}}$ and $T^{t}_{\text{P, P}}$, are calculated according to the transition matrix requirement of a Markov chain: the summation of a row in a transition matrix should be equal to one,

$$T^{t}_{\text{NBC, NBC}} = 1 - T^{t}_{\text{NBC, DOBC}} - T^{t}_{\text{NBC, P}}, \tag{4.5a}$$

$$T^{t}_{\text{P, P}} = 1 - T^{t}_{\text{P, DOBC}} - T^{t}_{\text{P, C}}. \tag{4.5b}$$

Table 4.4 summarizes the estimation for the one-step transition probability matrix for each age group given the normal observation.

Table 4.4: Transition matrices, T^t, given the normal observation (z^t =N) for each age group $\alpha(t)$, derived according to the sources in Table 4.3

age $\in [25,29]$ ($\alpha(t)=1$)

$$\begin{bmatrix} 0.9993749 & 0.0000817 & 0 & 0.0005434 & 0 \\ 0 & 0.7911233 & 0.2083333 & 0.0005434 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

age $\in [30,34]$ ($\alpha(t)=2$)

$$\begin{bmatrix} 0.9990366 & 0.0002634 & 0 & 0.0007000 & 0 \\ 0 & 0.7909667 & 0.2083333 & 0.0007000 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

age $\in [35,39]$ ($\alpha(t)=3$)

$$\begin{bmatrix} 0.9983675 & 0.0006028 & 0 & 0.0010297 & 0 \\ 0 & 0.7906370 & 0.2083333 & 0.0010297 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

age $\in [40,44]$ ($\alpha(t)=4$)

$$\begin{bmatrix} 0.9972098 & 0.0012093 & 0 & 0.0015808 & 0 \\ 0 & 0.7900858 & 0.2083333 & 0.0015808 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

age $\in [45,49]$ ($\alpha(t)=5$)

$$\begin{bmatrix} 0.9957612 & 0.0018829 & 0 & 0.0023559 & 0 \\ 0 & 0.7893107 & 0.2083333 & 0.0023559 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

age $\in [50,54]$ ($\alpha(t)=6$)

$$\begin{bmatrix} 0.9942273 & 0.0023360 & 0 & 0.0034367 & 0 \\ 0 & 0.7361466 & 0.2604167 & 0.0034367 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

age $\in [55,59]$ ($\alpha(t)=7$)

$$\begin{bmatrix} 0.9918927 & 0.0029255 & 0 & 0.0051819 & 0 \\ 0 & 0.7811429 & 0.2136752 & 0.0051819 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

age $\in [60,64]$ ($\alpha(t)=8$)

$$\begin{bmatrix} 0.9881242 & 0.0035850 & 0 & 0.0082908 & 0 \\ 0 & 0.8565741 & 0.1351351 & 0.0082908 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

age $\in [65,69]$ ($\alpha(t)=9$)

$$\begin{bmatrix} 0.9827372 & 0.0041190 & 0 & 0.0131439 & 0 \\ 0 & 0.8517210 & 0.1351351 & 0.0131439 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

age $\in [70,74]$ ($\alpha(t)=10$)

$$\begin{bmatrix} 0.9747491 & 0.0042474 & 0 & 0.0210034 & 0 \\ 0 & 0.8539966 & 0.1250000 & 0.0210034 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Table 4.4: (continued)

age $\in [75,79]$ ($\alpha(t) = 11$)

$$\begin{bmatrix} 0.9614493 & 0.0044760 & 0 & 0.0340747 & 0 \\ 0 & 0.8409253 & 0.1250000 & 0.0340747 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

age $\in [80,84]$ ($\alpha(t) = 12$)

$$\begin{bmatrix} 0.9385566 & 0.0042844 & 0 & 0.0571590 & 0 \\ 0 & 0.8178410 & 0.1250000 & 0.0571590 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

age $\in [85,100]$ ($\alpha(t) \in \{13,14,15\}$)

$$\begin{bmatrix} 0.8639572 & 0.0035794 & 0 & 0.1324633 & 0 \\ 0 & 0.7425367 & 0.1250000 & 0.1324633 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Specificity and sensitivity are the common measures to describe the accuracy (performance) of a medical test for disease detecting. The formula,

$$O^t_{\text{NBC, Ab}} = 1 - \text{specificity} \qquad \text{for } t \in \mathbb{T}. \tag{4.6}$$

is used to estimate $O^t_{\text{NBC, Ab}}$. Similarly, we use the sensitivity of mammography to yield the estimate

$$O^t_{\text{P, Ab}} = \text{sensitivity for } t \in \mathbb{T}. \tag{4.7}$$

Yankaskas et al. [52] report the sensitivity and the specificity of the screening mammography for women younger than 40 years. Their study provides the two parameters, $O^t_{\text{NBC, Ab}}$ and $O^t_{\text{P, Ab}}$, for the age groups 25 to 29, 30 to 34, and 35 to 39. For women older than age 40, we use the mammography performance data from the Breast Cancer Surveillance Consortium (BCSC) under the National Cancer Institute (NCI) [15]. The oldest age group from BCSC data is for women aged 75 to 89. We assume that the mammography performance stays the same for women aged 75 or older. Data from Yankaskas et al. [52] combined with the data from the BCSC [15] forms the complete observation matrix O^t. Table 4.5 summarizes the specificity and the sensitivity for each age group.

Table 4.5: Specificity and sensitivity by age group derived according to the sources in Table 4.3

Age group α	Age	Specificity(%)	Sensitivity(%)
			Preclinical
1	25–29	83.0	66.7
2	30–34	85.8	81.5
3	35–39	87.5	76.1
4	40–44	88.2	73.6
5	45–49	89.0	80.3
6	50–54	90.5	82.4
7	55–59	91.5	84.6
8	60–64	91.9	84.9
9	65–69	92.3	84.6
10	70–74	92.9	84.7
11	75–79	93.4	86.6
12	80–84	93.4	86.6
13–15	85–100	93.4	86.6

Given that we are solving small inverse problems, the lifetime mortality probabilities at the end of the time "subhorizon" of each age group is required for calculation. Zhang et al. [55] report the estimation of mortality probabilities by age and stage, which is in Table 4.6. For the last two age groups in our model, data from the Carolina Mammography Registry [12] is used to calculate the five-year and the ten-year mortality probabilities for women age 85 and above. The five-year mortality probability serves as the end-of-horizon value for the last age group, age 95 – 100, while the ten-year mortality probability serves as the-end-of horizon value for the age group 90–95. The description of the calculation of the end-of-horizon values can be found in [56].

Table 4.6: Lifetime mortality probability by age and stage

Age Group α	Age	In situ breast cancer	invasive breast cancer
1	25–29	0.13657	0.68533
2	30–34	0.13657	0.63111
3	35–39	0.13657	0.58458
4	40–44	0.12520	0.53218
5	45–49	0.12169	0.49751
6	50–54	0.11102	0.45382
7	55–59	0.11102	0.41585
8	60–64	0.10475	0.36916
9	65–69	0.10475	0.32890
10	70–74	0.10116	0.29506
11	75–79	0.09845	0.24037
12	80–84	0.09210	0.18407
13	≥ 85	0.08200	0.11600

4.4 Results and Discussion

Assuming a patient would like to minimize her lifetime breast cancer mortality probability, a screening policy which includes mammograms means that $R^t_{\mathrm{C,\,Ab}}$, the lifetime breast cancer mortality probability associated with being detected in state C, should be at least as large as $R^t_{\mathrm{P,\,Ab}}$, the lifetime breast cancer mortality probability associated with the state P, so that we have

$$R^t_{\mathrm{P,\,Ab}} \leqslant R^t_{\mathrm{C,\,Ab}} \text{ for all } t \in \mathbb{T}. \tag{4.8a}$$

The relation Eq. (4.8a) is consistent with all the available medical evidence; and if the inverse algorithm for a designated policy $\varphi \in \Phi$ yields the opposite result

$$R^t_{\mathrm{P,\,Ab}} > R^t_{\mathrm{C,\,Ab}} \text{ for some } t \in \mathbb{T}, \tag{4.8b}$$

then we may conclude that φ is not an operationally feasible policy.

For the policy of no screening at all decision epochs, so that

$$\varphi_{\mathrm{NS}}(\pi^t, t) \equiv \mathrm{NS} \text{ for all } t, \tag{4.9}$$

the inverse algorithm yields $R^t_{\mathrm{P,\,Ab}} = 1 > R^t_{\mathrm{C,\,Ab}} = 0$ for all t. We believe this is related to our method for addressing the degeneracy issue. In order to resolve the degeneracy issue, the inverse algorithm seeks the imputed rewards that maximize the summation of the differences between the value function of each alternative policy and the value function of the designated policy as formulated in Eq. (3.9).

In this breast cancer screening application with φ_{NS} as the designated policy, the imputed reward $R^t_{i,\mathrm{Ab}}$ represents the lifetime mortality probability associated with being detected in a

cancer state i at decision epoch t that returns (a) the largest summation of the differences between the value function of each alternative policy and the value function of the designated policy; and (b) the smallest value function for the designated policy. As we will discuss further in Chapter 5, the inverse algorithm selects the corner point ($R^t_{\text{P, Ab}} = 1, R^t_{\text{C, Ab}} = 0$) from the large set of combinations of the imputed rewards for the designated policy to resolve the degeneracy issue even though the imputed rewards of the designated policy are operationally infeasible. Different methods for resolving the degeneracy issue may need to be explored to address the issue of the operational infeasibility of the imputed rewards. Similar reasoning applies to the policy $\varphi_{\text{NS@10}} \equiv [\text{NS, NS, NS, NS, NS, NS, NS, NS, NS, S}]$ as well, where the imputed rewards for this policy are $R^{\alpha(t)}_{\text{P, Ab}} > 0$ and $R^{\alpha(t)}_{\text{C, Ab}} = 0$ for each age group, $\alpha(t) \in \{2, 4, 6, 8, 11, 13\}$.

Figure 4.2, through Figure 4.7 contain plots of the following: (a) the average imputed rewards $\overline{R}^{\alpha(t)}_{i,\text{Ab}}$ for $i \in \{\text{P, C}\}$ and age group $\alpha(t)$ for all policies, where

$$\overline{R}^{\alpha(t)}_{i,\text{Ab}}(\widetilde{\varphi}) = \frac{1}{\mathcal{L}} \sum_{w=1}^{\mathcal{L}} \left\{ \frac{1}{|\mathscr{G}|} \sum_{\pi^0 \in \mathscr{G}} R^{\mathcal{L}(\alpha(t)-1)+w}_{i,\text{Ab}}(\widetilde{\varphi}) \right\} \text{ for } i \in \{\text{P, C}\}; \tag{4.10}$$

where $R^w_{i,\text{Ab}}(\widetilde{\varphi})$ for $i \in \{\text{P, C}\}$ is the imputed reward associated with being detected in state P or C of the designated policy at decision epoch w; and (b) the average expected number of mammograms over the remaining time horizon $\{0, 1, \ldots, H\}$ for the designated policy $\widetilde{\varphi}$ at decision epoch 0,

$$\overline{M}^0(\widetilde{\varphi}) = \frac{1}{|\mathscr{G}|} \sum_{\pi^0 \in \mathscr{G}} M^0(\widetilde{\varphi}; \pi^0) \text{ for all } \widetilde{\varphi} \in \Phi. \tag{4.11}$$

Table 4.7 summarizes the following: (a) the imputed rewards, $\overline{R}^{\alpha(t)}_{i,\text{Ab}}$, for $i \in \{\text{P, C}\}$; and (b) the average expected number of mammograms over the remaining time horizon $\{0, 1, \ldots, H\}$ for the designated policy $\widetilde{\varphi}$ at decision epoch 0, $\overline{M}^0(\widetilde{\varphi})$. In the following figures, we single out

for special attention the imputed results for two breast cancer screening guideline that have received widespread recognition:

- The American Cancer Society (ACS) [44] recommends annual screening for a woman of age 40+ so long as she is in good health; and
- The US Preventive Services Task Force (USPSTF) [50] recommends biennial screening from age 50 to age 74.

In the following discussion, let φ_{ACS} and φ_{USP} denote the breast cancer screening policies of ACS and USPSTF, respectively.

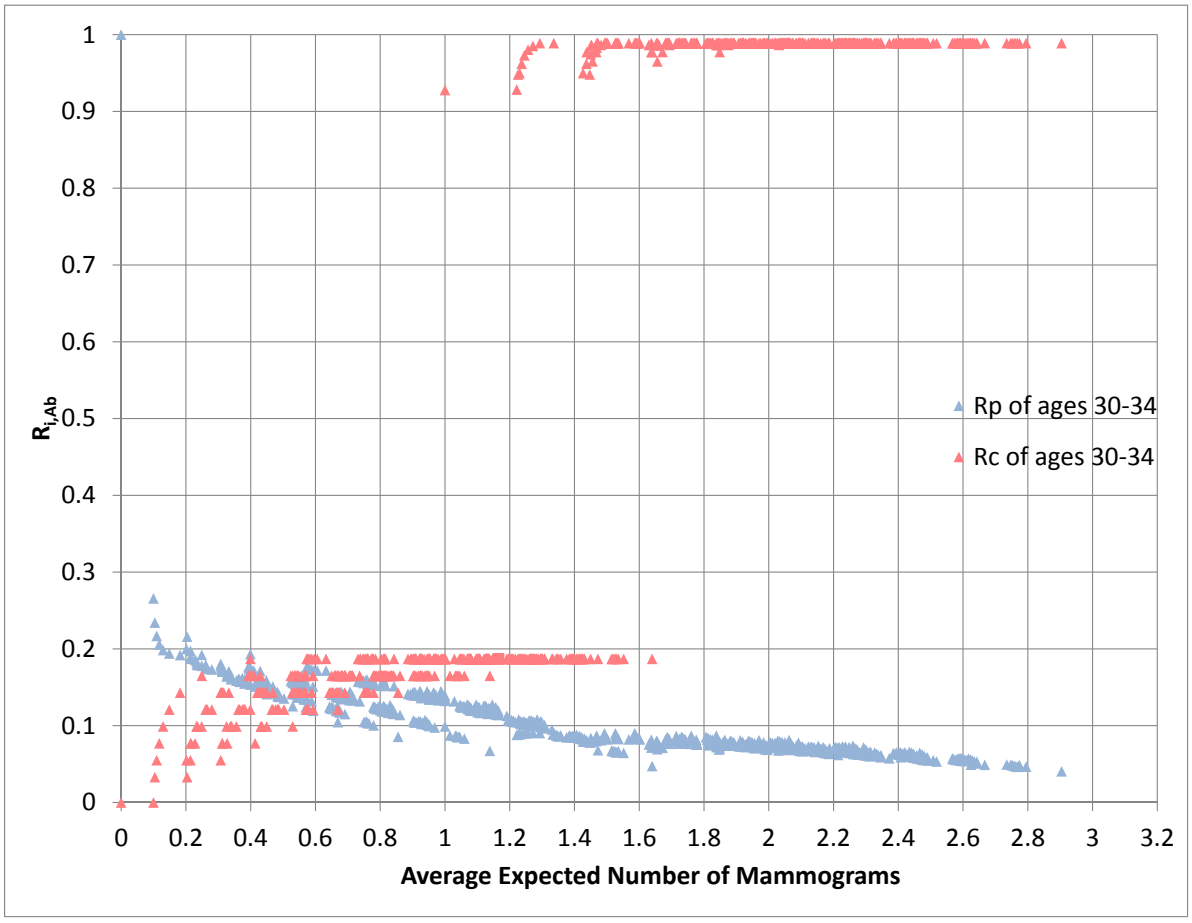


Figure 4.2: The average imputed rewards $\overline{R}_{i,\mathrm{Ab}}^{\alpha(t)}(\widetilde{\varphi})$ for $i \in \{\mathrm{P}, \mathrm{C}\}$ and the average expected number of mammograms $\overline{M}^{0}(\widetilde{\varphi})$ for the age group 30–34, $\alpha(t) = 2$.

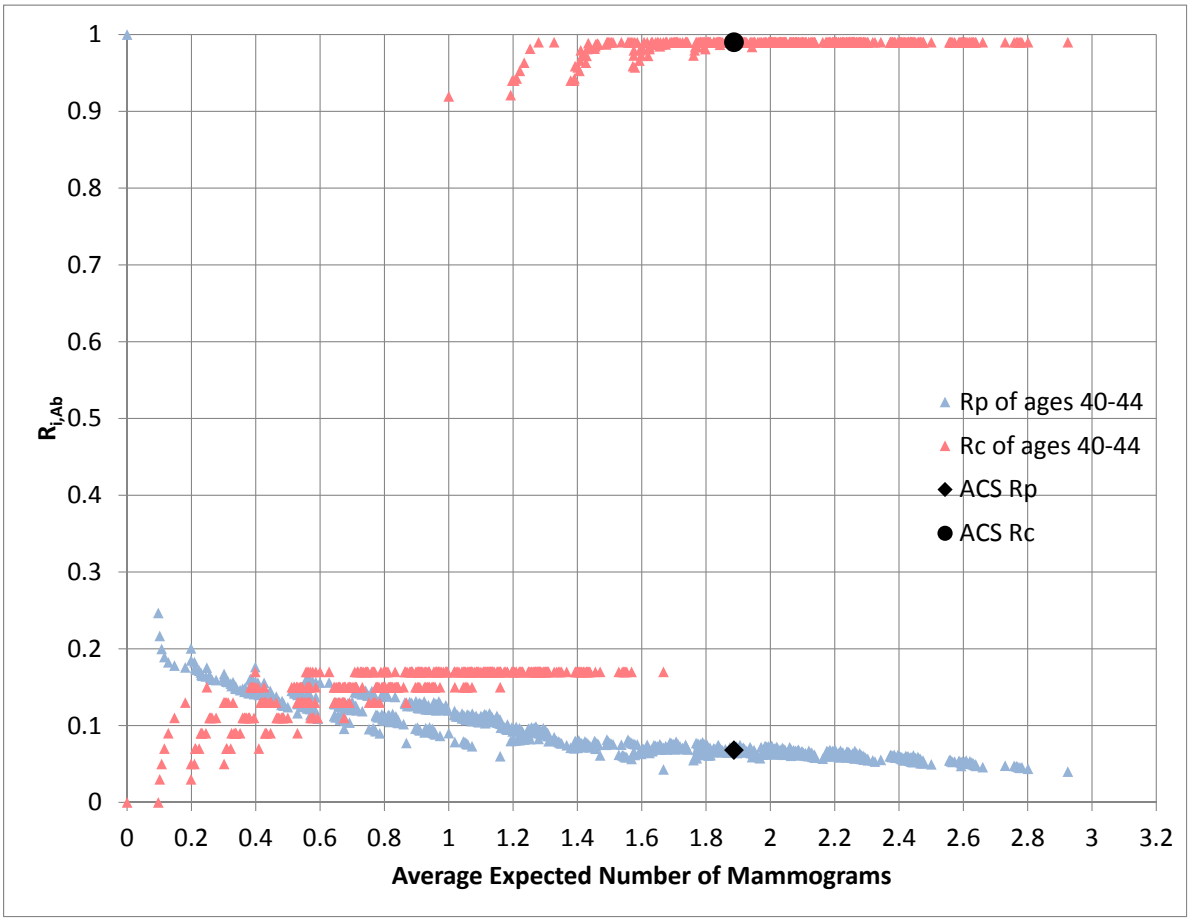


Figure 4.3: The average imputed rewards $\overline{R}^{\alpha(t)}_{i,\mathrm{Ab}}(\widetilde{\varphi})$ for $i \in \{\mathrm{P}, \mathrm{C}\}$ and the average expected number of mammograms $\overline{M}^{0}(\widetilde{\varphi})$ for the age group 40–44, $\alpha(t) = 4$. The black circle and the black diamond correspond to the average imputed rewards for the ACS screening policy φ_{ACS}.

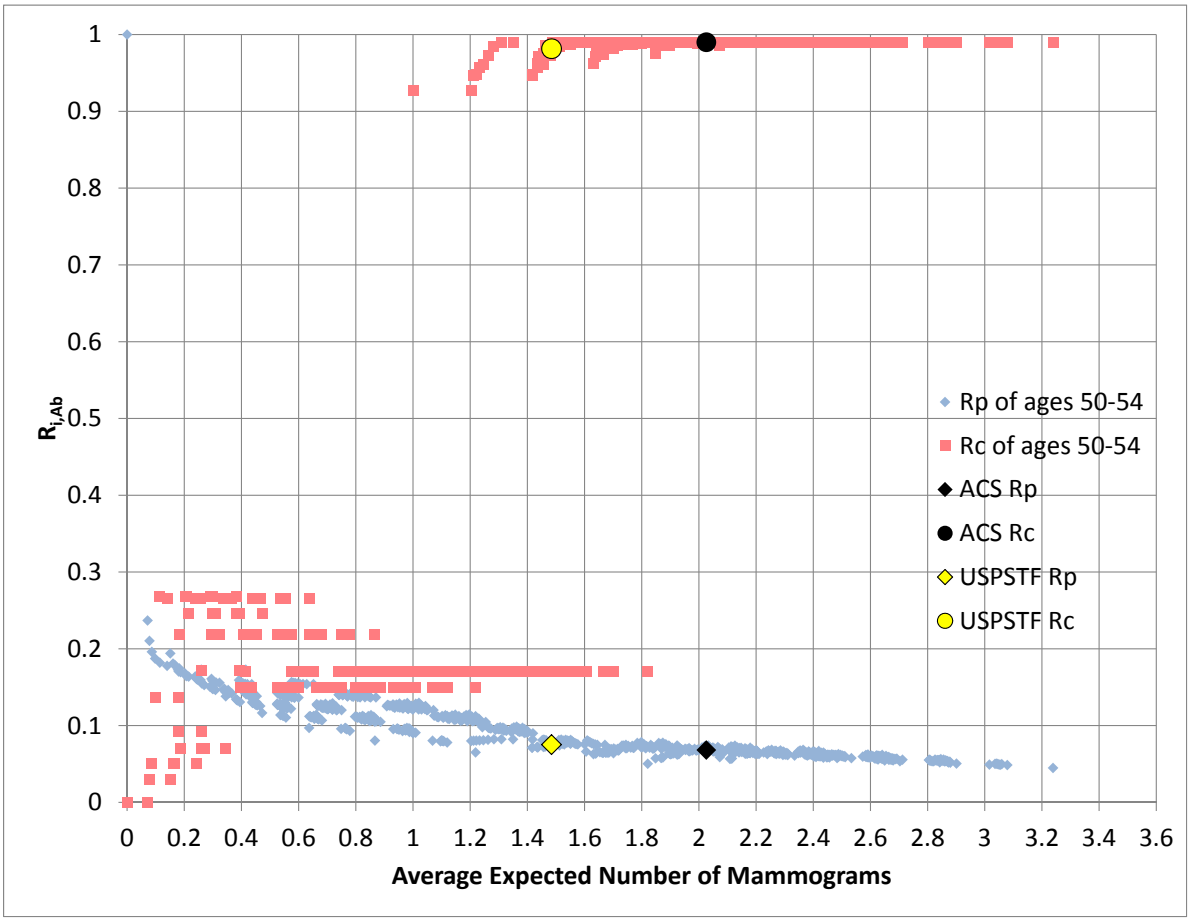


Figure 4.4: The average imputed rewards $\overline{R}_{i,\mathrm{Ab}}^{\alpha(t)}(\widetilde{\varphi})$ for $i \in \{\mathrm{P}, \mathrm{C}\}$ and the average expected number of mammograms $\overline{M}^{0}(\widetilde{\varphi})$ for the age group 50–54, $\alpha(t) = 6$. The black circle and the black diamond correspond to the average imputed rewards for the ACS screening policy φ_{ACS}. The yellow circle and the yellow diamond correspond to the USPSTF policy φ_{USP}.

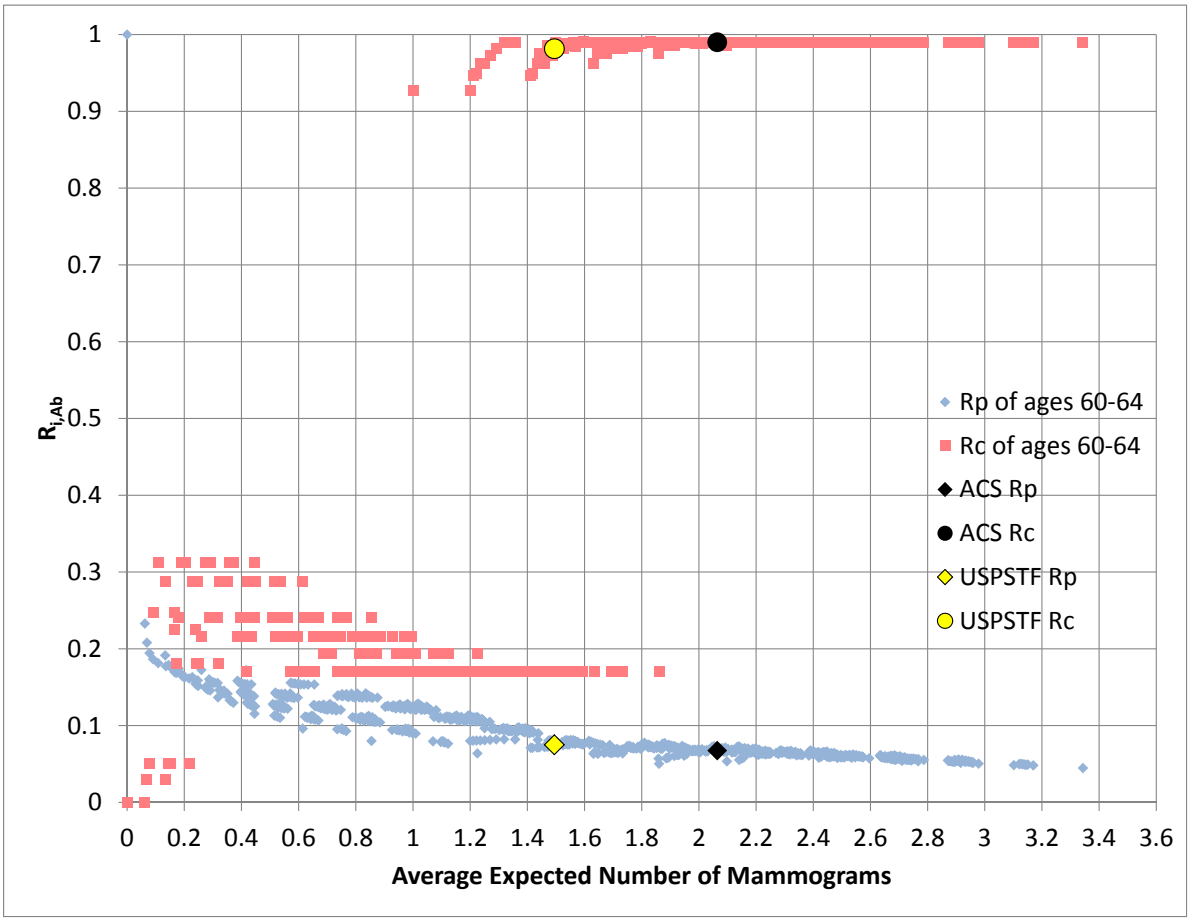


Figure 4.5: The average imputed rewards $\overline{R}_{i,\mathrm{Ab}}^{\alpha(t)}(\widetilde{\varphi})$ for $i \in \{\mathrm{P}, \mathrm{C}\}$ and the average expected number of mammograms $\overline{M}^{0}(\widetilde{\varphi})$ for the age group 60–64, $\alpha(t) = 8$. The black circle and the black diamond correspond to the average imputed rewards for the ACS screening policy φ_{ACS}. The yellow circle and the yellow diamond correspond to the USPSTF policy φ_{USP}.

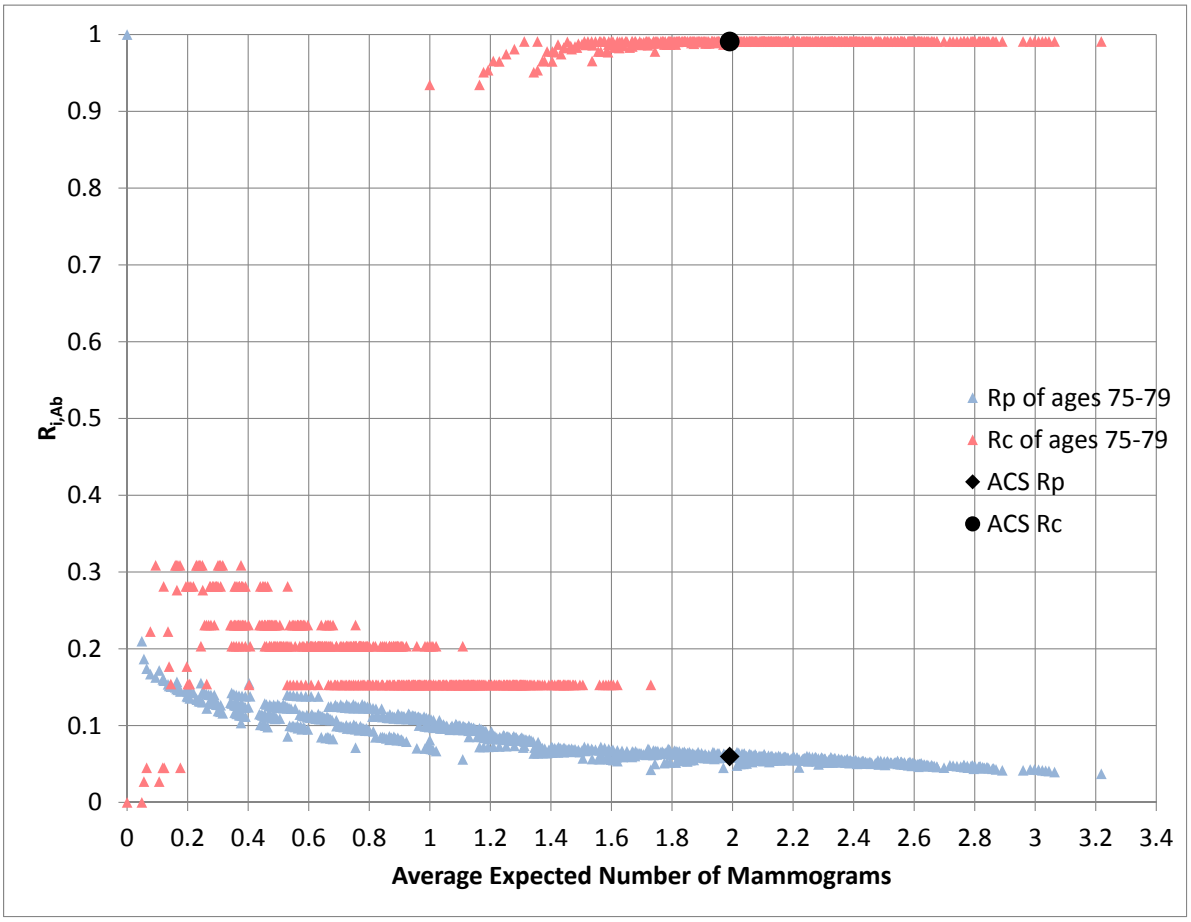


Figure 4.6: The average imputed rewards $\overline{R}^{\alpha(t)}_{i,\mathrm{Ab}}(\widetilde{\varphi})$ for $i \in \{\mathrm{P}, \mathrm{C}\}$ and the average expected number of mammograms $\overline{M}^{0}(\widetilde{\varphi})$ for the age group 75–79, $\alpha(t) = 11$. The black circle and the black diamond correspond to the average imputed rewards for the ACS screening policy φ_{ACS}.

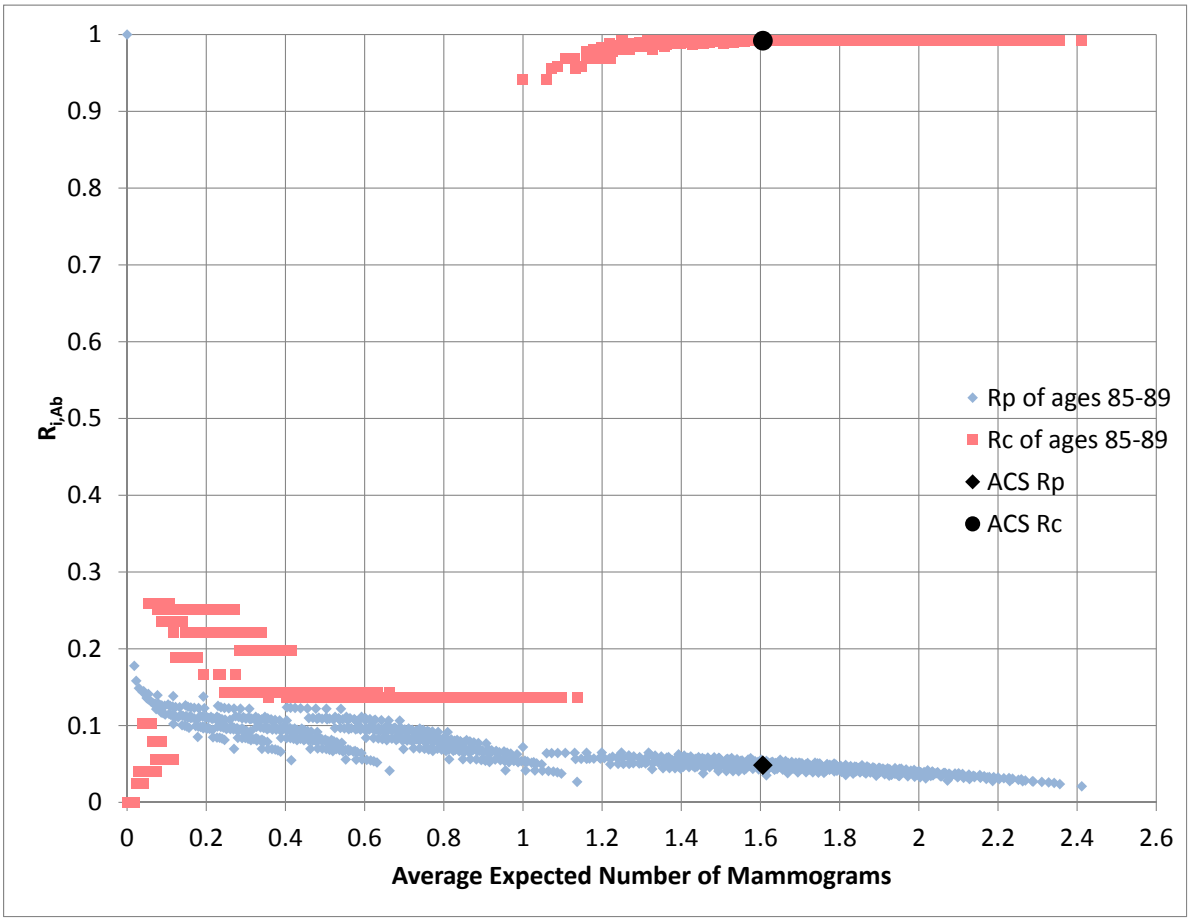


Figure 4.7: The average imputed rewards $\overline{R}^{\alpha(t)}_{i,\mathrm{Ab}}(\widetilde{\varphi})$ for $i \in \{\mathrm{P}, \mathrm{C}\}$ and the average expected number of mammograms $\overline{M}^{0}(\widetilde{\varphi})$ for the age group 85–89, $\alpha(t) = 13$. The black circle and the black diamond correspond to the average imputed rewards for the ACS screening policy φ_{ACS}.

Table 4.7: The imputed rewards $\overline{R}^{\alpha(t)}_{i,\mathrm{AB}}(\widetilde{\varphi})$ for $i \in \{\mathrm{P}, \mathrm{C}\}$ and the expected number of mammograms $\overline{M}^{0}(\widetilde{\varphi})$ for the policies recommended by the ACS and the USPSTF.

Age Group α	Age		Designated Policy $\widetilde{\varphi} =$	
			φ_{USP}	φ_{ACS}
4	40–44	$\overline{M}^{0}(\widetilde{\varphi})$		1.88694
		$\overline{R}^{\alpha(t)}_{\mathrm{P,\,Ab}}(\widetilde{\varphi})$		0.06798
		$\overline{R}^{\alpha(t)}_{\mathrm{C,\,Ab}}(\widetilde{\varphi})$		0.99000
6	50–54	$\overline{M}^{0}(\widetilde{\varphi})$	1.48431	2.02605
		$\overline{R}^{\alpha(t)}_{\mathrm{P,\,Ab}}(\widetilde{\varphi})$	0.07527	0.06809
		$\overline{R}^{\alpha(t)}_{\mathrm{C,\,Ab}}(\widetilde{\varphi})$	0.98167	0.99000
8	60–64	$\overline{M}^{0}(\widetilde{\varphi})$	1.49446	2.06430
		$\overline{R}^{\alpha(t)}_{\mathrm{P,\,Ab}}(\widetilde{\varphi})$	0.07504	0.06749
		$\overline{R}^{\alpha(t)}_{\mathrm{C,\,Ab}}(\widetilde{\varphi})$	0.98167	0.99000
11	75–79	$\overline{M}^{0}(\widetilde{\varphi})$		1.99068
		$\overline{R}^{\alpha(t)}_{\mathrm{P,\,Ab}}(\widetilde{\varphi})$		0.05973
		$\overline{R}^{\alpha(t)}_{\mathrm{C,\,Ab}}(\widetilde{\varphi})$		0.99100
13	85–89	$\overline{M}^{0}(\widetilde{\varphi})$		1.60615
		$\overline{R}^{\alpha(t)}_{\mathrm{P,\,Ab}}(\widetilde{\varphi})$		0.04840
		$\overline{R}^{\alpha(t)}_{\mathrm{C,\,Ab}}(\widetilde{\varphi})$		0.99200

These graphs illustrate the relationship between the imputed rewards associated with both states and the expected number of mammograms. When the expected number of mammograms increases, the imputed reward associated with state P decreases and the imputed reward associated with state C increases. This relationship is consistent with our hypothesis: the more screening tests that are scheduled in a policy, the higher the breast cancer mortality probability associated with state C must be and the lower the corresponding breast cancer mortality probability associated with state P must be. This relationship shows what we mentioned earlier: more screening tests, which translate to earlier detections, must result in higher survival probabilities in order for the lower the mortality probabilities associated with state P to be optimal. However, when taking account of the effort associated with executing the screening policy, we can see that the slope of the imputed reward associated with state P is decreasing as the expected number of mammograms increases. The decreasing slope may suggest that an extra screening test does not decrease the imputed reward associated with state P as much as the expected number of mammograms increases. Hence, this observation suggests the existence of a trade-off threshold between the expected number of mammograms and the imputed rewards.

Also, we notice that the imputed reward associated with state C forms two major clusters that distinguish the policies clearly: (i) the policies with a low imputed reward associated with state C and less effort required to execute the designated policy; and (ii) the policies with a high imputed reward associated with state C and more effort required to execute the designated policy. The feature of the policies in cluster (i) is a screening test at the first decision epoch so that the expected number of mammograms for these policies is greater than one. These policies with high imputed rewards associated with state C tend to have mammograms earlier in the time horizon because the disease has a high mortality probability, i.e., when the imputed rewards are high, it is better that a patient would like to perform a mammogram as soon as possible, typically at the first decision epoch.

Figure 4.8 and Figure 4.9 show all the policies for each of the different age groups as a function of (a) the average value function over the grid $\mathscr{G}$,

$$\overline{V}^0(\widetilde{\varphi}) = \frac{1}{|\mathscr{G}|} \sum_{\pi^0 \in \mathscr{G}} \widetilde{V}^0(\pi^0, \widetilde{\varphi}) \text{ for all } \widetilde{\varphi} \in \Phi, \tag{4.12}$$

where $\widetilde{V}^0(\pi^0, \widetilde{\varphi})$ is evaluated with the imputed rewards, $R_{i,\text{Ab}}^{\alpha(t)}, i \in \{\text{P, C}\}$ associated with that designated policy $\widetilde{\varphi}$; and (b) the average expected number of mammograms over the remaining time horizon $\{0, 1, \ldots, H\}$ for the designated policy $\widetilde{\varphi}$ at decision epoch 0. Each dot in the graphs represents a policy, i.e., there are 1024 points for each age group shown in the graphs.

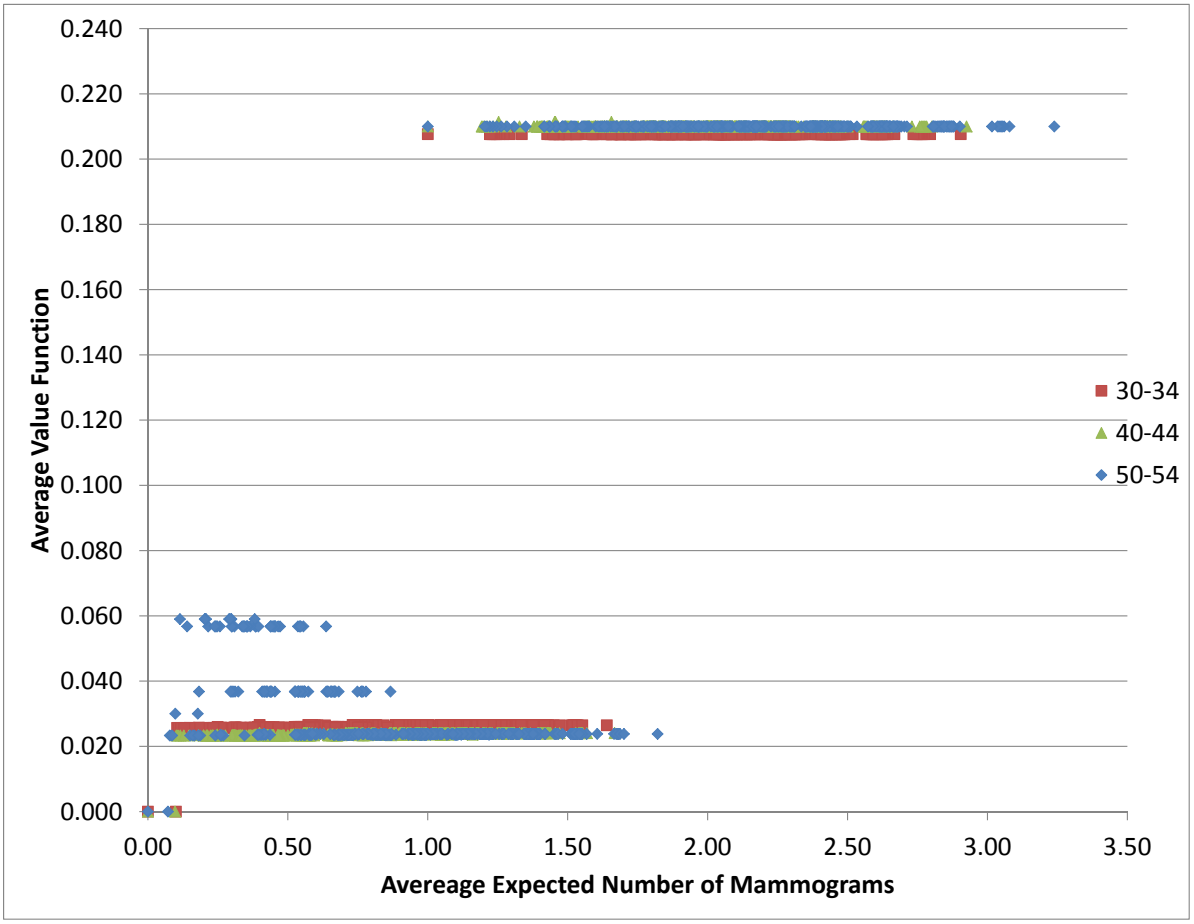


Figure 4.8: The average value function $\overline{V}^0(\widetilde{\varphi})$ and the average expected number of mammograms $\overline{M}^0(\widetilde{\varphi})$ for the age groups, 30–34, 40–44, and 50–54.

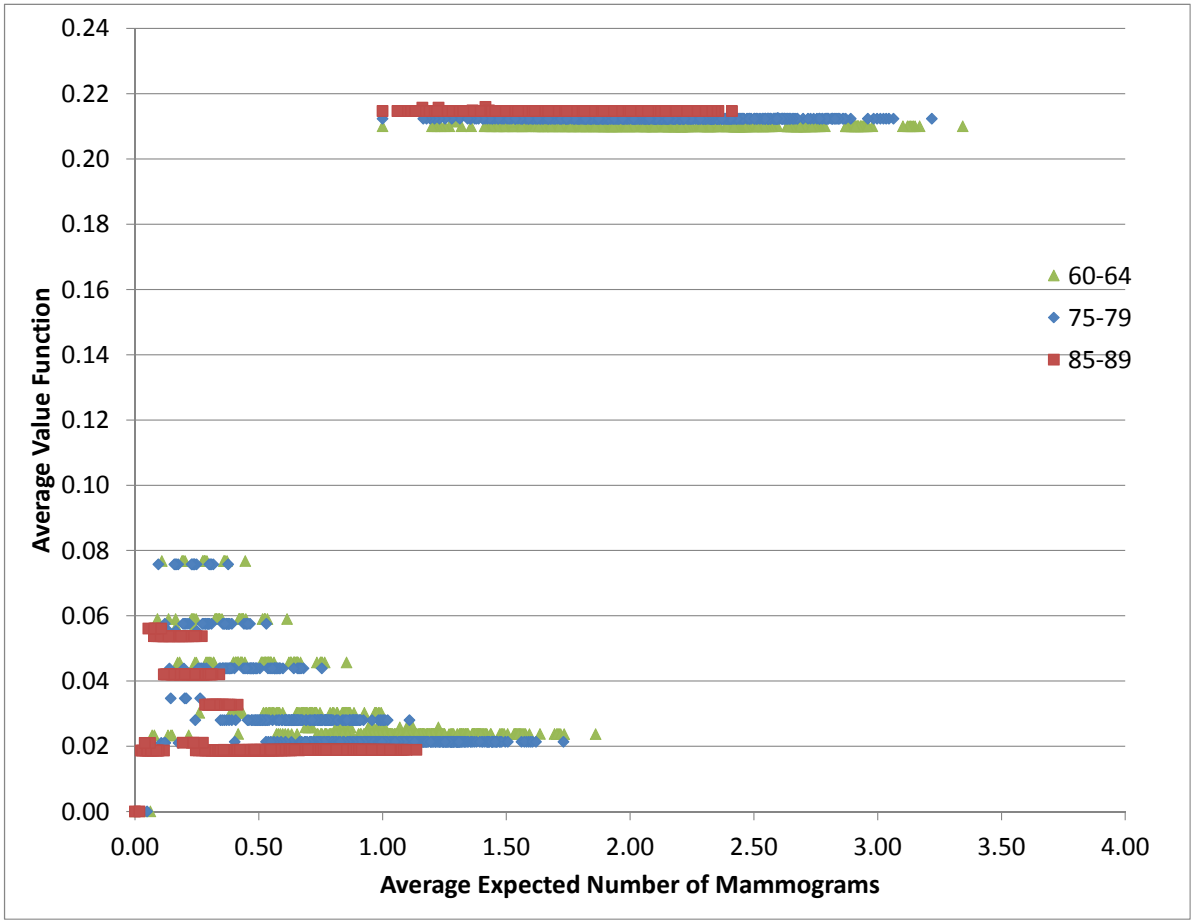


Figure 4.9: The average value function $\overline{V}^0(\widetilde{\varphi})$ and the average expected number of mammograms $\overline{M}^0(\widetilde{\varphi})$ for the age groups, 60–64, 75–79, and 85–89.

Similar to the clusters of the imputed reward associated with state C shown in Figure 4.2, Figure 4.3, Figure 4.4, Figure 4.5, Figure 4.6, and Figure 4.7, we can see two major clusters are in both graphs that distinguish the policies clearly: (i) policies which result in lower overall breast cancer mortality probabilities with less effort required to execute the designated policy; and (ii) policies that result in higher overall breast cancer mortality probability with more effort required to execute the designated policy. The common feature in those policies with a larger value function is a screening test at the first decision epoch so that the expected number of mammograms for these policies is greater than one. The result of these policies with the high value function result and the high expected number of mammograms is consistent with the reasons for the imputed rewards. Also, in each of the clusters, the value functions of most of the policies are very close to each other, which may suggest that a policy with an effective screening test schedule (i.e., lower expected number of mammograms) can reduce breast cancer mortality probabilities as much as a policy with a frequent screening test schedule (i.e., higher expected number of mammograms).

In Figure 4.8 and Figure 4.9, the policies which are located on the x-axis are (i) the policy with no screening test at all decision epochs, φ_{NS}; and (ii) the policy with one screening test at the last decision epoch, $\varphi_{\mathrm{NS@10}}$. Although these policies requires very little effort to execute, the imputed rewards are operationally infeasible, i.e, $R^t_{\mathrm{P,\,Ab}} > R^t_{\mathrm{C,\,Ab}}$. Hence, these policies may not be recommended to all age groups.

The numerical results suggest that the initial belief state, $\pi_x = [1,0,0,0,0]$, is a special initial belief state for all age groups because, given the optimality assumption for some designated policies, the inverse problem is never feasible even without the robust equality constraint, Eq. (3.9e). Based on a detailed investigation of the corresponding LP problems, we found that the infeasibility of such an LP problem comes from the constraint, Eq. (3.9b), where the designated policy, $\widetilde{\varphi}$, which includes no screening at the first decision epoch $t = 0$, has to be at

least equivalent to or outperform the alternative policy, $\widehat{\varphi}$, which includes a similar screening pattern of scheduled screenings but with a screening test at the first decision epoch. In other words, the alternative policy, $\widehat{\varphi}$, which is identical to $\widetilde{\varphi}$ except for a screening test at the first decision epoch, is better than the designated policy $\widetilde{\varphi}$, which does not include a screening test at the first decision epoch. Evaluating $\widetilde{\varphi}$ and $\widehat{\varphi}$ solely on the basis of mortality probabilities may suggest that these policies cannot be recommended because no imputed rewards exist for which the designated policy can outperform all alternatives. However, when we take the effort of executing a policy into account, the alternative policy $\widehat{\varphi}$ that is identical to $\widetilde{\varphi}$ except for a screening test at the first decision epoch requires more effort to execute compared with $\widetilde{\varphi}$. Hence, future research will explore this issue further.

The tolerance, ε, of each age group relaxes the strong equality constraint Eq. (2.45a) and allows all inverse problems for any designated time-dependent policy to be feasible. The tolerance ε varies across age groups as shown in Table 4.8. The younger the age group is, the larger the tolerance has to be so that the inverse problem is feasible for every $\widetilde{\varphi} \in \Phi$. This result supports the hypothesis that the younger a woman is, the more uncertainty regarding her health in the future.

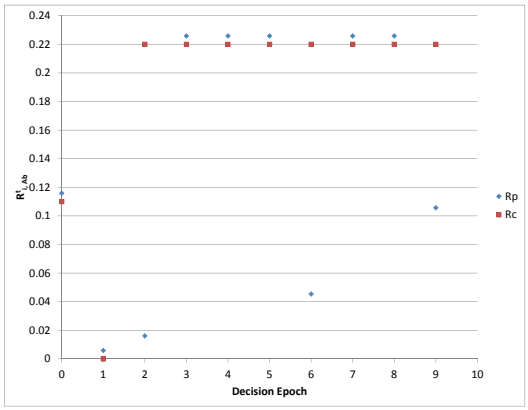
Table 4.8: The tolerance, ε of each testing age group

Age Group α	Age	Tolerance, $\max(\varepsilon_\varphi)$ for $\varphi \in \Phi$
2	30–34	0.11
4	40–44	0.10
6	50–54	0.10
11	75–79	0.09
13	85–89	0.08

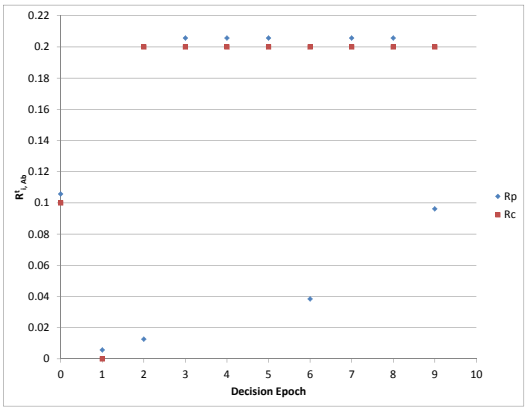
The tolerance as shown in Table 4.8 for each age group is the tolerance which allows

all inverse problems for any designated time-dependent policy to be feasible. However, the tolerances of two different selections for the designated policy $\widetilde{\varphi}$ may be different from each other. Also, for a designated policy $\widetilde{\varphi}$, the sum of the base value $R^{\tau}_{i,\mathrm{Ab}}$ for $i \in \{\mathrm{P}, \mathrm{C}\}$ at decision epoch $t \in \mathbb{T}$ and the tolerance ε is the upper bound on the imputed reward $R^{\beta}_{i,\mathrm{Ab}}$ at decision epoch β for $\beta \in \{\tau+1, \tau+2, \ldots, \tau+\mathcal{L}-1\}$ and $i \in \{\mathrm{P}, \mathrm{C}\}$ as given in Eq. (3.9e). For example, for the age group 75–79, the base value and the upper bound on the mortality probability $R^{\alpha(t)}_{\mathrm{P,\,Ab}}$ of the policy, $\varphi = [\mathrm{NS,S,S,NS,NS,NS,S,NS,NS,S}]$, are 0.0951 and 0.1851, respectively. Further investigation is required to identify more accurate (smaller) tolerances to improve estimate of the imputed rewards. One possible method is to solve a minimization problem involving the tolerance ε, or to provide the tolerance, $\varepsilon(\varphi, \pi^0)$ by the combination of a policy $\varphi \in \Phi$ and a initial belief state $\pi^0 \in \mathscr{G}$.

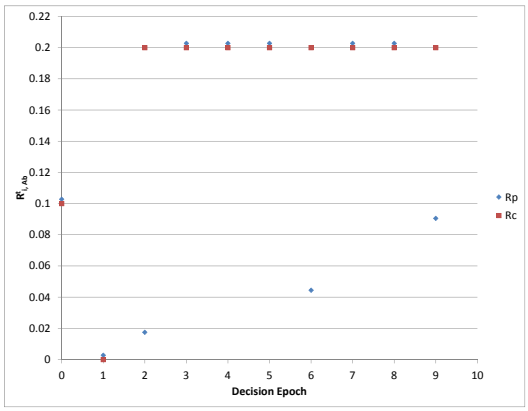
Through the time horizon, when a screening test is taken, the mortality probabilities of a designated policy may drop. For example, the mortality probabilities of the designated policy $\varphi = [\mathrm{NS,S,S,NS,NS,NS,S,NS,NS,S}]$ for all testing age groups are shown in Figure 4.10. This particular policy φ includes total four mammograms which are taken at the decision epochs $t = \{1,3,6,9\}$, respectively. In Figure 4.10, the mortality probability associated with being detected in state P for each testing age group, $R^{t}_{\mathrm{P,\,Ab}}$ for $t = \{0,1,\ldots,9\}$ fluctuates according to the timing of the mammograms.



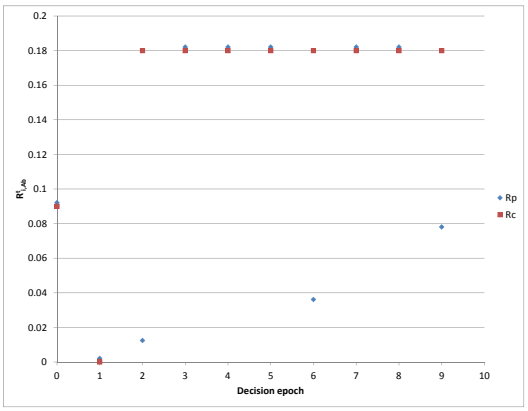
(a) age group 30–34



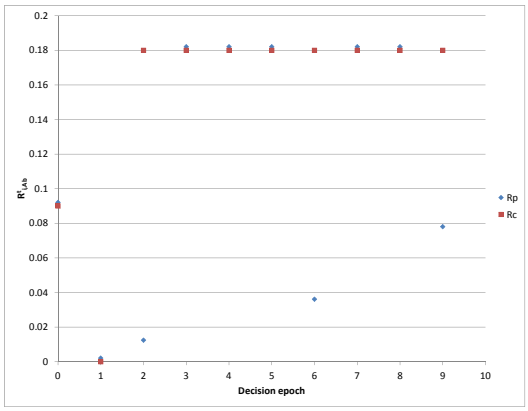
(b) age group 40–44



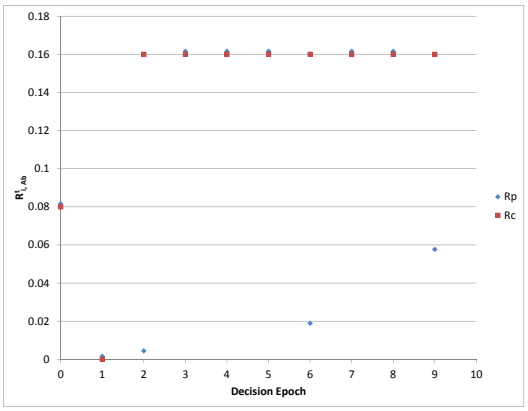
(c) age group 50–54



(d) age group 60–64



(e) age group 75–79



(f) age group 85–89

Figure 4.10: $R^t_{i,\mathrm{Ab}}$ for $i \in \{\mathrm{P}, \mathrm{C}\}$ and $t \in \{0, 1, \ldots, 9\}$ of the policy,φ = [NS, S, S, NS, NS, NS, S, NS, NS, S], for all testing age groups.

We will explore possible alternative methods for addressing the fluctuation in the imputed rewards by capturing the relationship between the imputed rewards at different decision epochs, such as applying the exponential smoothing method [9] to the imputed rewards over the time horizon so that the imputed rewards increase along with a patient's age for this breast cancer screening policy application.

As suggested by Figure 4.2 through Figure 4.9, the result from solving several small problems for each age group without a connection between each age group cannot tell a complete story about the breast cancer mortality probability for a policy which considers a patient from age 25 to age 100. For example, the impact on the starting age of recommended mammograms for the two major policies recommended by the ACS and the USPSTF (age 40 and 50, respectively) does not show in these graphs. The USPSTF policy recommends no screening between ages 40 and 49. However, the imputed rewards for the do-nothing policy in age groups 40–44 are 1 and 0 for state P and C, similar to all other testing age groups, as shown in Figure 4.3. Hence, the next step must be to connect the results for all age groups together into a comprehensive analysis of the breast cancer mortality probabilities that are imputed from different time-dependent screening policies.

4.5 Conclusion

In this chapter, we describe in detail the execution of the inverse POMDP breast cancer screening model of Section 3.4, including: (a) the discussion of the method for selecting a grid $\mathscr{G}$ of initial belief states in Section 4.1; (b) a formulation of the method for evaluating the imputed reward result in Section 4.2; and (c) a description of all the data sources in Section 4.3. The numerical results are presented in Section 4.4.

The numerical results show the relationship between the imputed rewards associated with

being detected in different states, which provides a general guide for selecting a screening policy. The numerical results suggest that the timing of a screening test has a significant impact on the breast cancer mortality associated with being detected in a particular cancer stage at a particular age, so that the overall lifetime mortality probability is affected as well. A trade-off exists between the mortality probabilities associated with the cancer states and the effort of executing a policy. A limitation of our numerical analysis is due to the fact that we take the average of the value function and the expected number of mammograms over the grid $\mathscr{G}$, which may not be the best method for aggregating all results over the state space. A possible alternative method would be a weighted average by using the breast cancer risk factor calculator to assign different weights to each grid point $\pi_x \in \mathscr{G}$.

With regard to the initial belief state, $\pi_x = [1,0,0,0,0]$ represents a special case because of the infeasibility of some designated policies whose common feature is no screening test at the first decision epoch. For the infeasible inverse problem of such designated policies, the infeasibility is caused by the constraint that the designated policy $\widetilde{\varphi}$ with no screening at decision epoch $t = 1$ has to be better than or at least equal to the alternative policy $\widehat{\varphi}$ that has a screening test at the first decision epoch. Further study on this special case is necessary to understand and solve the infeasibility issue.

To search for an insightful answer, an equality constraint with tolerance is added into the inverse problem. The analysis of the inverse problems for different age groups indicates that the tolerance varies with the age group. The younger the age group is, the more variable the tolerance is. Also, the tolerance varies across the designated policies that are chosen for analysis. Hence, more research is required to understand the behavior of the imputed rewards given different tolerances. Also, one may explore different methods, such as exponential smoothing, to capture the relationship of the imputed rewards at different decision epochs.

Chapter 5

Validation and Sensitivity Analysis

In Section 5.1, we first conduct the validation of the algorithm to solve the inverse POMDP for breast cancer screening with time-dependent screening policies. The second part of this chapter covers the sensitivity analysis. In Section 5.2, we analyze the sensitivity of the results to random variations of as much as $\pm 10\%$ in the one-step transition probabilities $\{T_{i,j}^{\ t}(a,\ell) : i, j \in \Omega; a \in \mathcal{A}; t \in \mathbb{T}; \ell \in \mathcal{Z}\}$ and in the observation probabilities $\{O_{i,\ell}^{t}(a) : i \in \Omega; t \in \mathbb{T}; \ell \in \mathcal{Z}; a \in \mathcal{A}\}$.

5.1 Validation

The purpose of this validation section is to confirm that the imputed rewards which are obtained from the inverse algorithm do make the designated policy the optimal policy. The validation method is to solve the forward POMDP given different sets of rewards. In order to maintain the time-dependency of the screening policies, the solution method for the forward POMDP is the policy evaluation method. For a given initial belief state, π_x, the policy evaluation method first calculates the sample path in the belief state space Π and then calculates the value function result with the given rewards for each time-dependent policy. Let $\breve{\varphi}$ denote the optimal policy

which is obtained from the policy evaluation method. The optimal policy $\breve{\varphi}$ returns the smallest value function result with less number of mammograms.

First, the equality constraint, Eq. (2.45a), without the tolerance ε is applied to the rewards over the time horizon. Hence, the number of unknowns (rewards) are reduced to two, $R^{\alpha(t)}_{\text{P, Ab}}$ and $R^{\alpha(t)}_{\text{C, Ab}}$, for each age group. In the two-dimensional reward space, we discretize each dimension with step length equal 0.01. Hence, a total of 10,201 combinations of the two rewards are used to solve the forward POMDP problem as specified in Eq. (3.5) and Eq. (3.6) of Section 3.3 given the constraint, $0 \leqslant R^{\alpha(t)}_{i,\text{Ab}} \leqslant 1$ for $i \in \{\text{P, C}\}$. Two age groups, ages 50–54 and 75–79, are selected to conduct the validation with the equality constraint Eq. (2.45a), and three different initial belief states, [0.8, 0.2, 0, 0, 0], [0.6, 0.2, 0.2, 0, 0], and [0.3, 0.7, 0, 0, 0], are selected to perform this validation. Figure 5.1 to Figure 5.6 show the optimal policy for each reward combination for a particular initial belief state.

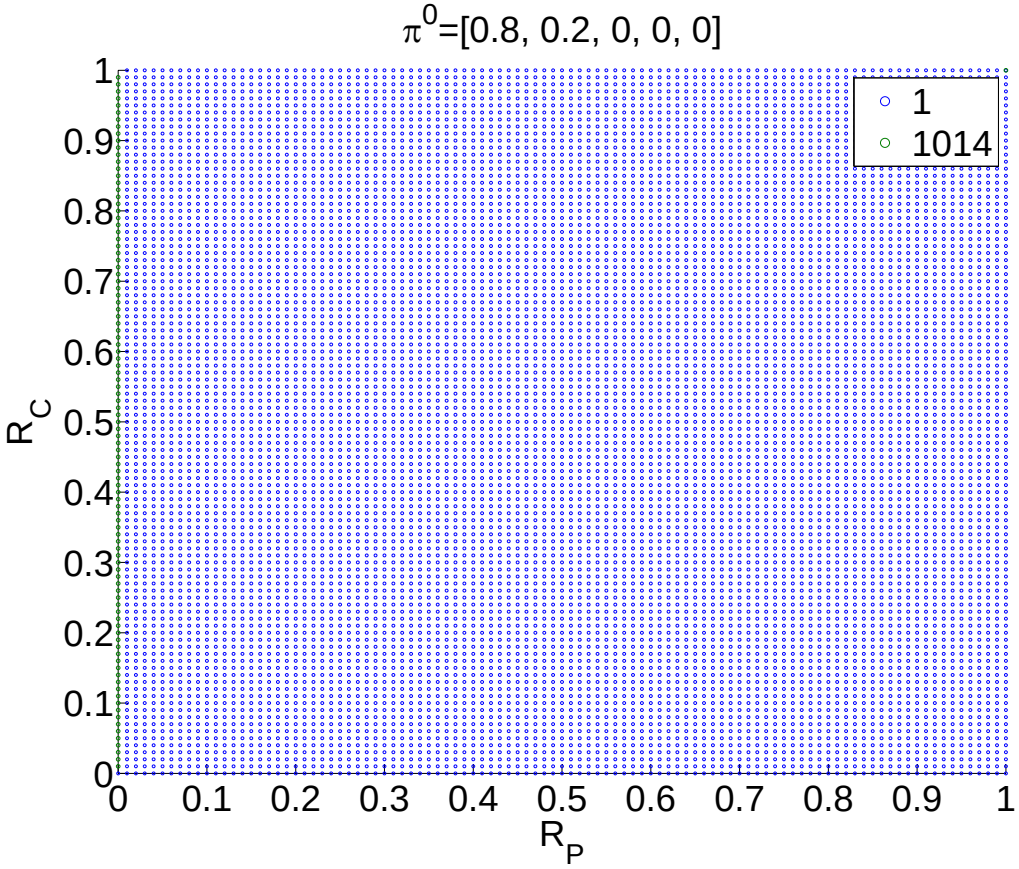


Figure 5.1: The optimal policies of the forward POMDP for different initial belief state π^0 =[0.8, 0.2, 0, 0, 0] given different reward combinations solved with the policy evaluation method for the age group 50–54. The numbers in the legend represent the numbering of the policies where "1" represents the policy φ_{NS} and "1014" represents the policy with nine screening tests at the first nine decision epochs.

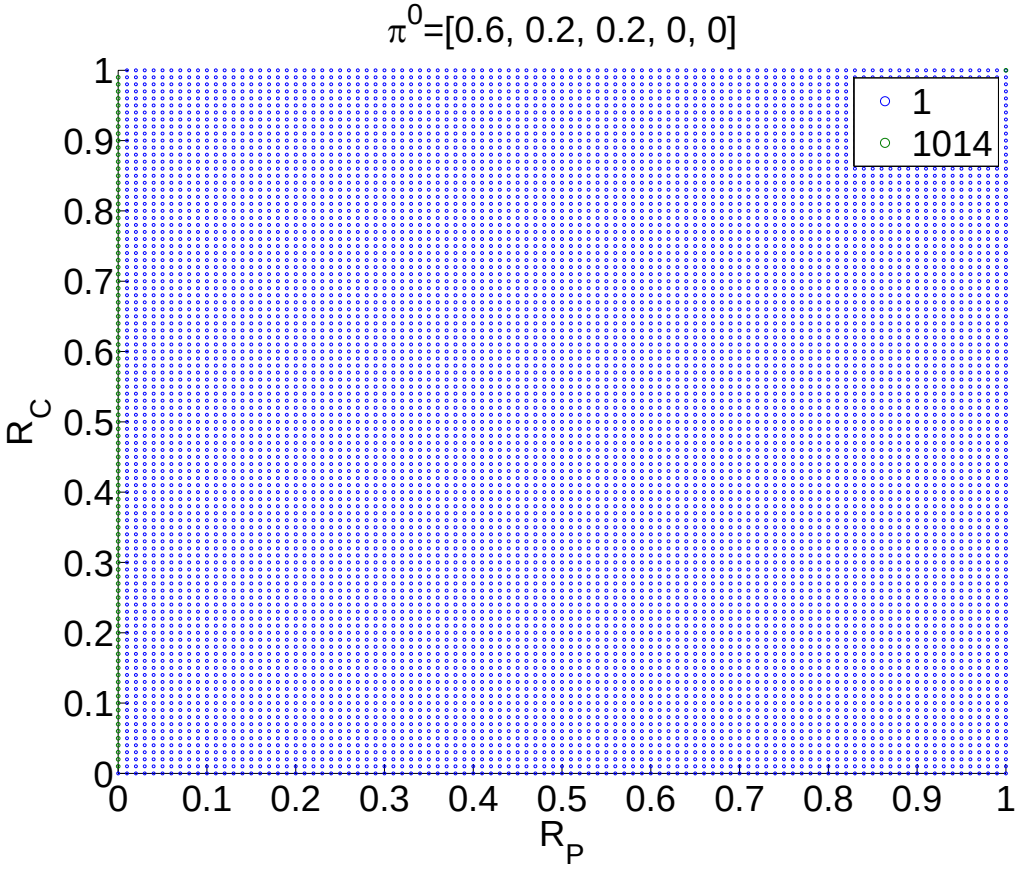


Figure 5.2: The optimal policies of the forward POMDP for different initial belief state π^0 =[0.6, 0.2, 0.2, 0, 0] given different reward combinations solved with the policy evaluation method for the age group 50–54. The numbers in the legend represent the numbering of the policies where "1" represents the policy φ_{NS} and "1014" represents the policy with nine screening tests at the first nine decision epochs.

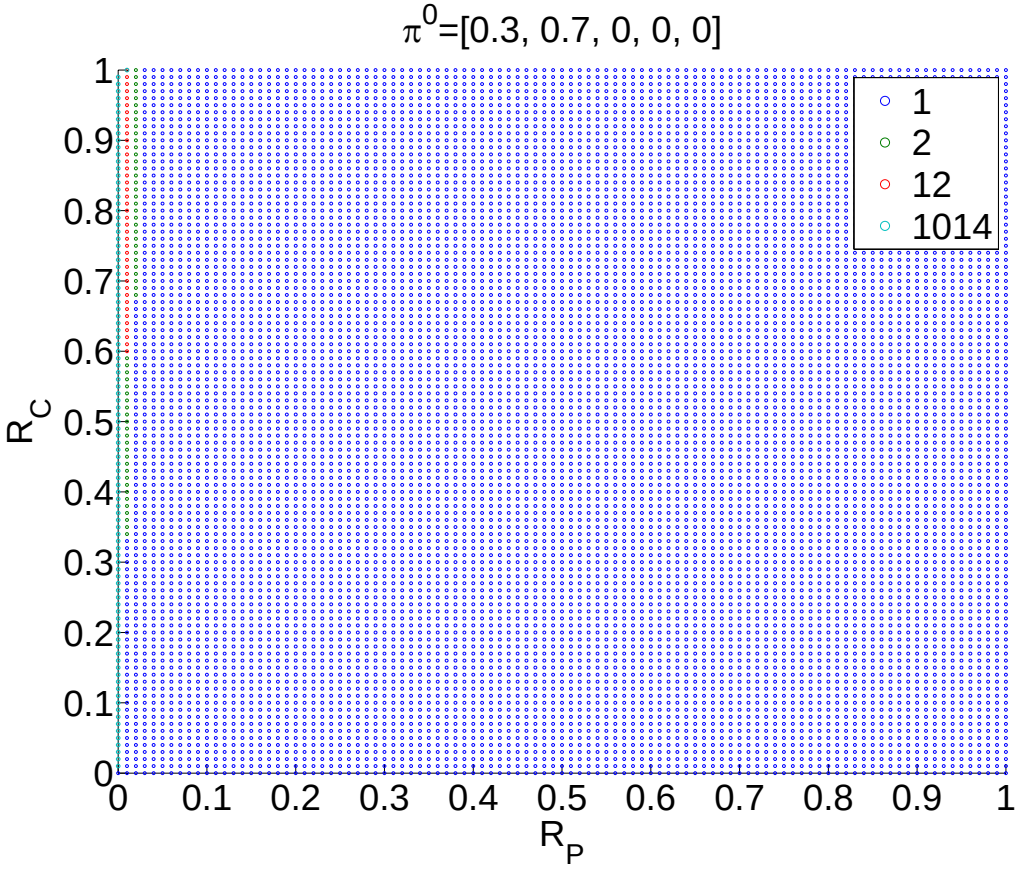


Figure 5.3: The optimal policies of the forward POMDP for different initial belief state π^0 =[0.3, 0.7, 0, 0, 0] given different reward combinations solved with the policy evaluation method for the age group 50–54. The numbers in the legend represent the numbering of the policies where "1" represents the policy φ_{NS}; "2" represents the policy with one screening at $t = 1$; "12" represents the policy with two screenings at $t = 1$ and $t = 2$, respectively; and "1014" represents the policy with nine screening tests at the first nine decision epochs.

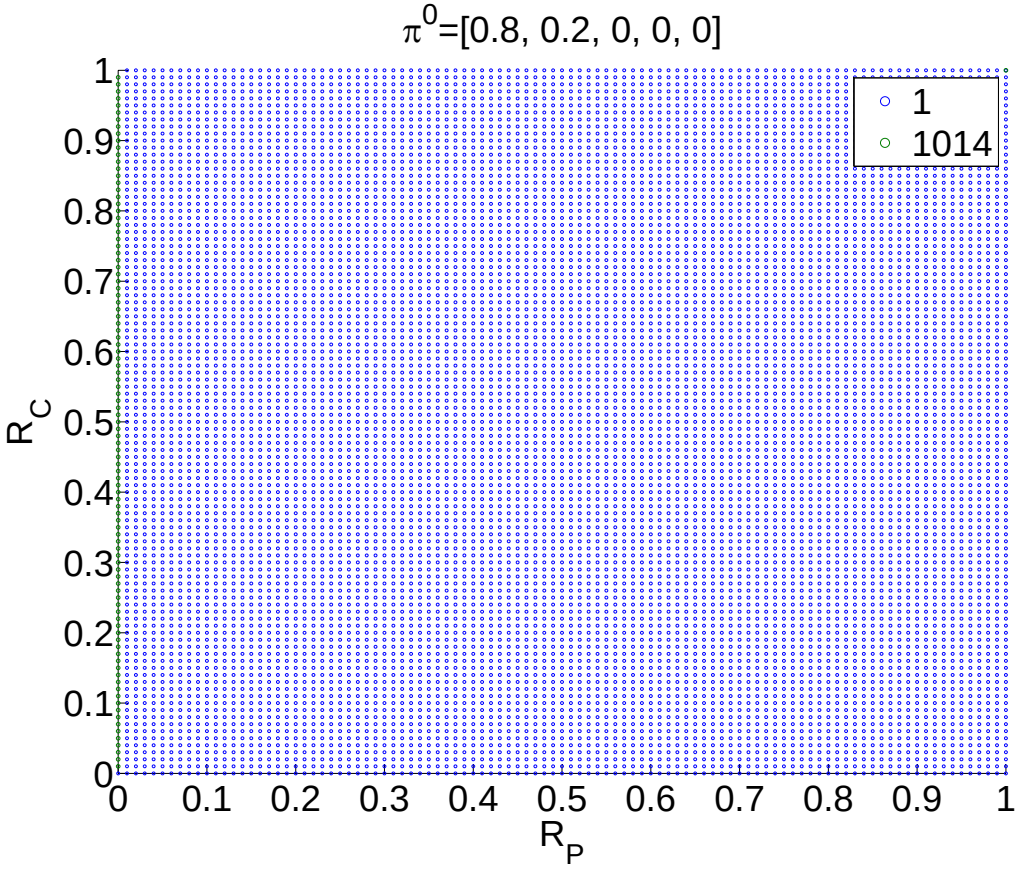


Figure 5.4: The optimal policies of the forward POMDP for different initial belief state π^0 =[0.8, 0.2, 0, 0, 0] given different reward combinations solved with the policy evaluation method for the age group 75–79. The numbers in the legend represent the numbering of the policies where "1" represents the policy φ_{NS} and "1014" represents the policy with nine screening tests at the first nine decision epochs.

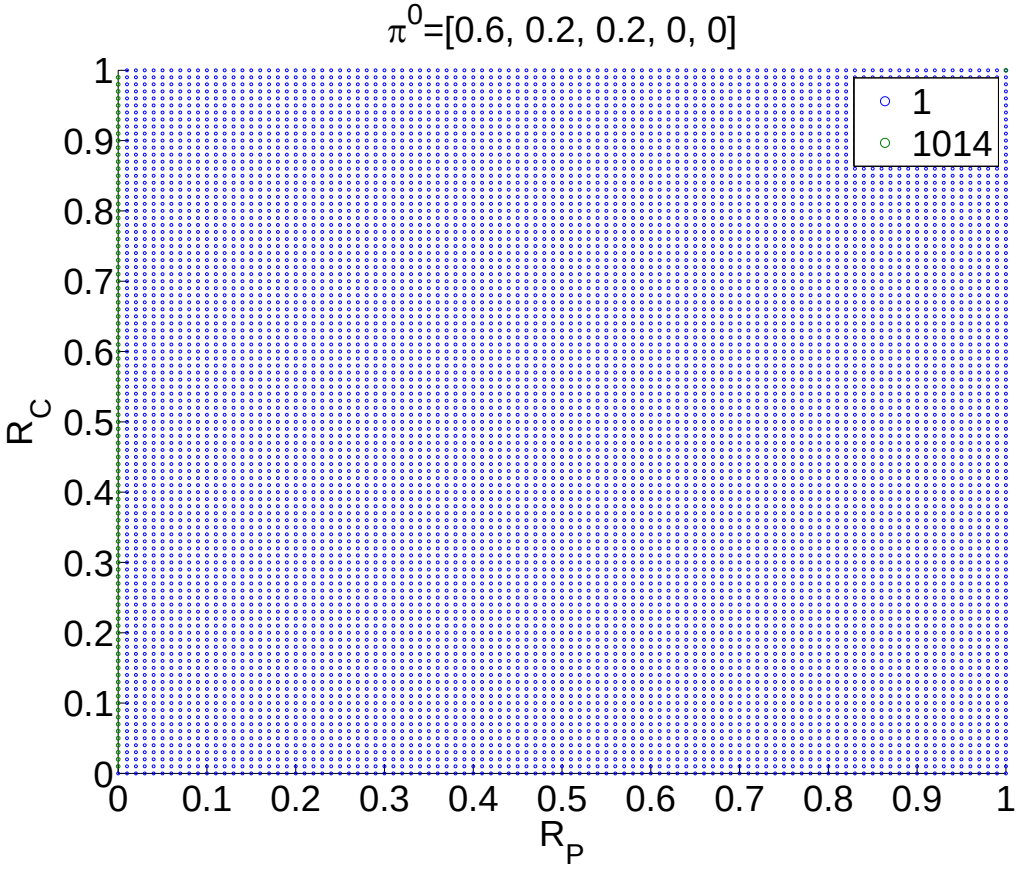


Figure 5.5: The optimal policies of the forward POMDP for different initial belief state π^0 =[0.6, 0.2, 0.2, 0, 0] given different reward combinations solved with the policy evaluation method for the age group 75–79. The numbers in the legend represent the numbering of the policies where "1" represents the policy φ_{NS} and "1014" represents the policy with nine screening tests at the first nine decision epochs.

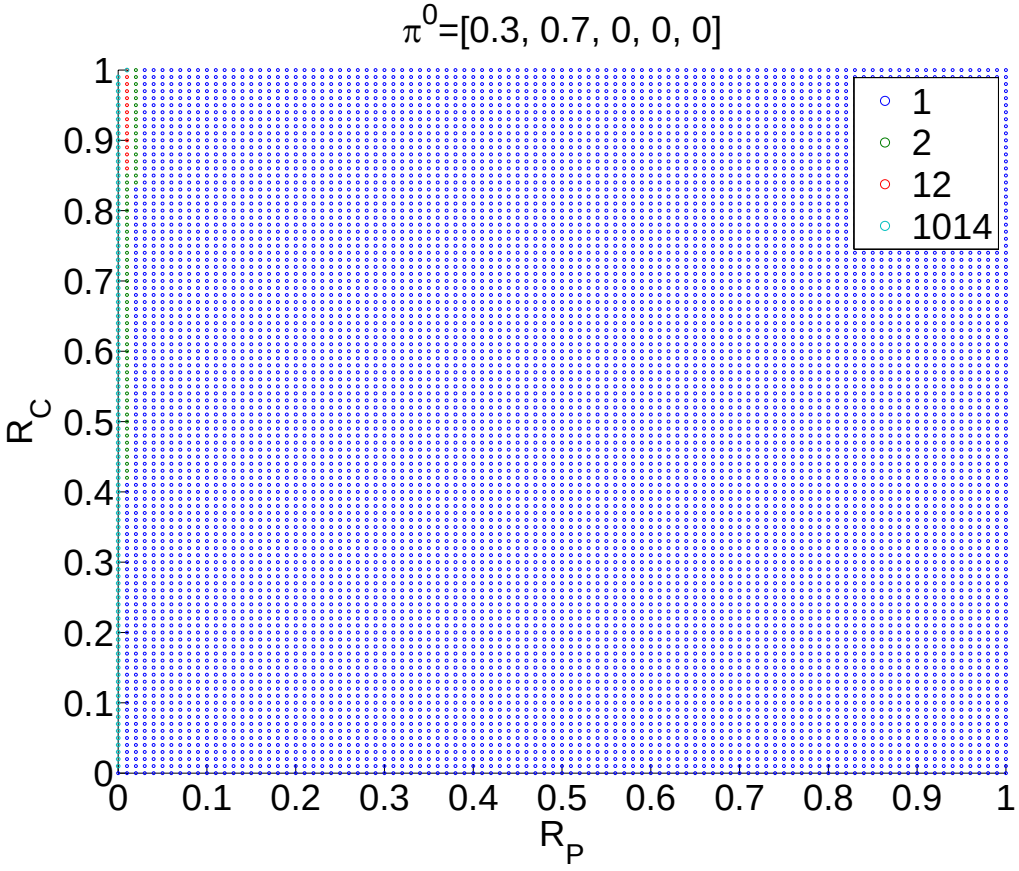


Figure 5.6: The optimal policies of the forward POMDP for different initial belief state π^0 =[0.3, 0.7, 0, 0, 0] given different reward combinations solved with the policy evaluation method for the age group 75–79. The numbers in the legend represent the numbering of the policies where "1" represents the policy φ_{NS}; "2" represents the policy with one screening at $t = 1$; "12" represents the policy with two screenings at $t = 1$ and $t = 2$, respectively; and "1014" represents the policy with nine screening tests at the first nine decision epochs.

As shown in Figure 5.1 to Figure 5.6, the majority of the two-dimensional reward space, the blue area, is dominated by the do-nothing policy φ_{NS} for the three initial belief states and both age groups. For the combinations of rewards such that the mortality probability associated with being detected in state P, $R^{\alpha(t)}_{\text{P, Ab}}$, is close to zero but not equal to zero (the area that is not in blue), the optimal policy is different for different initial belief states. The optimal policies in this small colorful area are a function of the initial belief states and the combinations of the rewards. In terms of a patient's health, when the belief state π_x describing a patient's health becomes worse for a given combination of $R^{\alpha(t)}_{\text{P, Ab}}$ and $R^{\alpha(t)}_{\text{C, Ab}}$, the optimal policy requires more screening tests at the beginning of the time horizon. Hence, we summarize the pattern of the colorful area: (a) for a given initial belief state π_x, the closer the combination of mortality probabilities is to the point (0, 1), the more screening tests will be required near the beginning of the time horizon; (b) for a given combination of mortality probabilities, the higher the probability of being in state P, a screening test is more helpful, which is consistent with the purpose of a screening test (i.e., early detection); and (c) when the mortality probability combination is located on the y-axis so that $R^{\alpha(t)}_{\text{P, Ab}} = 0$, the optimal policy performs a screening test at each of the first nine decision epoches.

From the inverse problem, one can confirm that, for the do-nothing policy φ_{NS}, the imputed rewards, $R^{\alpha(t)}_{\text{P, Ab}} = 1$ and $R^{\alpha(t)}_{\text{C, Ab}} = 0$, are located at the corner point (i.e., (1, 0)) of the blue region. However, the imputed rewards for all other policies contain some noise because of the tolerance ε. In order to validate our inverse algorithm, the tolerance ε should be taken into account as well. Hence, the imputed rewards $R^{\alpha(t)}_{\text{P, Ab}}$ and $R^{\alpha(t)}_{\text{C, Ab}}$, which are computed by the inverse problem for each combination of the initial belief state and the designated policy, are used as the rewards to solve the forward problem with the policy evaluation method. A flowchart, Figure 5.7, shows the steps of validation process. Let $\check{V}^0(\pi_x, \check{\varphi})$ denote the value function of the policy $\check{\varphi}$. Age groups 40–44, 50–54 and 75–79 are selected to illustrate this

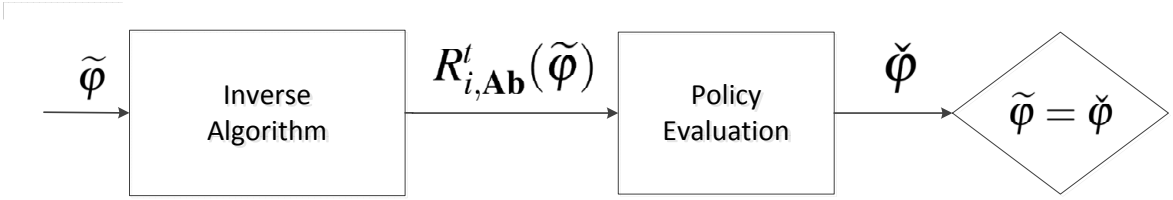


Figure 5.7: The flowchart of the validation procedure.

policy evaluation. From Figure 5.7, the first step is to selected a designated policy $\widetilde{\varphi}$. Then, the inverse algorithm uses the designated policy to calculate the imputed rewards $R^t_{i,\mathrm{Ab}}(\widetilde{\varphi})$. Next, the forward POMDP is solved as described in Section 3.3 by the policy evaluation method. The policy evaluation method returns the optimal policy $\check{\varphi}$. The last step is to compare the optimal policy which is obtained from the policy evaluation is equal to the designated policy $\widetilde{\varphi}$.

From the solution of the policy evaluation with the imputed rewards for a given initial belief state $\pi_x \in \mathscr{G}$ and a given designated policy $\widetilde{\varphi} \in \Phi$, the designated policy for the given initial belief state may not return the lowest value function result, i.e, it may happen that $\widetilde{V}^0(\pi_x, \widetilde{\varphi}) \geqslant \check{V}^0(\pi_x, \check{\varphi})$. However, the difference between the value function for the optimal policy identified using the policy evaluation and the value function for the designated policy, $\widetilde{V}^0(\pi_x, \widetilde{\varphi}) - \check{V}^0(\pi_x, \check{\varphi})$, is approximately less than or equal to 10^{-6}. The small difference suggests that the designated policy may be equivalent to the optimal policy that is obtained using the policy evaluation method. This suggests that the optimality condition for the designated policy (Eq. (2.40)) could be satisfied. Table 5.1 shows two examples for each testing age group, where the initial belief states are randomly selected from the grid $\mathscr{G}$, and we take the designated policy $\widetilde{\varphi}$ for the inverse POMDP to be one of the following: (a) φ_{ACS}, the screening policy recommended by ACS; or (b) φ_{USP}, the screening policy recommended by USPSTF.

Table 5.1: Representative examples that show the difference between the recommended policies, $\widetilde{\varphi}_{\text{ACS}}$ and $\widetilde{\varphi}_{\text{USPSTF}}$, and the optimal policy $\breve{\varphi}$, which is obtained from the policy evaluation of the forward POMDP using the imputed rewards $R_{i,\text{Ab}}^{\alpha(t)}$ for $i \in \{\text{P, C}\}$ which are obtained from the inverse POMDP algorithm.

Age Group α(Age)	Policy & Initial belief state π^0	Difference	$\overline{R}_{P,\text{Ab}}^{\alpha(t)}(\widetilde{\varphi})$	$\overline{R}_{C,\text{Ab}}^{\alpha(t)}(\widetilde{\varphi})$
4 (40–44)	$\widetilde{\varphi} =$ φ_{ACS}			
	$\breve{\varphi} =$ [S, NS, NS, NS, NS, NS, NS, NS, NS, NS]			
	$\pi^0 =$ [0.4,0.6,0,0,0]	2.1213e-07	0.06798	0.99
	$\widetilde{\varphi} =$ φ_{ACS}			
	$\breve{\varphi} =$ [S, NS, S, NS, NS, NS, NS, NS, NS, NS]			
	$\pi^0 =$ [0.2,0.7,0.1,0,0]	4.5733e-08	0.06798	0.99
6 (50–54)	$\widetilde{\varphi} =$ φ_{ACS}			
	$\breve{\varphi} =$ [S, NS, NS, NS, NS, NS, NS, NS, NS, NS]			
	$\pi^0 =$ [0.5,0.1,0.4,0,0]	6.8745e-07	0.06809	0.99
	$\widetilde{\varphi} =$ φ_{USP}			
	$\breve{\varphi} =$ [S, NS, NS, NS, NS, NS, NS, NS, NS, NS]			
	$\pi^0 =$ [0.4,0.5,0.1,0,0]	4.3234e-07	0.07527	0.98167
11 (75–79)	$\widetilde{\varphi} =$ φ_{ACS}			
	$\breve{\varphi} =$ [S, NS, NS, NS, NS, NS, NS, NS, NS, NS]			
	$\pi^0 =$ [0.9,0.1,0,0,0]	1.8118e-06	0.05973	0.9910
	$\widetilde{\varphi} =$ φ_{USP}			
	$\breve{\varphi} =$ φ_{USP}			
	$\pi^0 =$ [0.7,0.1,0.2,0,0]	0	0.06697	0.9825

5.2 Sensitivity Analysis

Two age groups are selected to conduct the sensitivity analysis. The two age groups are ages 40–44, and ages 75–79. The sampling method described at the beginning of this section is used to generate 50 different sets of transition matrices and observation matrices given that each element varies by at most $\pm 10\%$ from the original value. Also, only four different policies are selected to test that are:

Policy 1 No screening within the age group;

Policy 2 Annual screening as recommended by the American Cancer Society [44];

Policy 3 Biennial screening as recommended by the US Preventive Services Task Force [50];

Policy 4 Screening at each decision epoch.

Uniform Spacings Method

For the purpose of multidimensional sensitivity analysis, the uniform-spacings method to sample a new transition matrix is developed and presented in this section. Given a $\nu \times \nu$ Markov chain transition matrix T, the method treats each row in the transition matrix as a beginning point, $\widehat{\Omega}_i, i \in \{1, \ldots, \nu\}$, in the ν-dimensional Euclidean space and randomly samples a new point $\widetilde{\Omega}_i$ on the standard simplex in ν-dimensional Euclidean space,

$$\mathcal{H} \equiv \left\{ y = [y_1, \ldots, y_\nu] \in \mathbb{R}^\nu : 0 \leqslant y_i \leqslant 1 \text{ and } \sum_{i=1}^{\nu} y_i = 1 \right\}, \tag{5.1}$$

which is formed according to the Markov chain transition matrix property, i.e., the summation of all elements in each row is equal to one. The sampled point $\widetilde{\Omega}_i$ also has to be within the

uncertainty set Γ_i, which is assumed to consist of all points in $\mathcal{H}$ whose components fall within $\pm 10\%$ of the corresponding components of $\widehat{\Omega}_i$.

Let $\mathcal{W}$ denote the number of sampled transition matrices for which we would like to test the inverse algorithm, and let $\mathcal{X}$ denote the index of the $\mathcal{X}$-th sampled transition matrix for $\mathcal{X} = 1, \ldots, \mathcal{W}$. Therefore, the sampling algorithm generates $\mathcal{W}$ realizations (samples) of each row in the transition matrix, and yielding a total of $\mathcal{W} \times \nu$ rows and $\mathcal{W}$ realizations (samples) of the $\nu \times \nu$ transition matrix.

At the beginning of sampling a new point, the result from the previous iteration, $\widetilde{\Omega}_i(\mathcal{X}-1)$, serves as the new "center point" $\widetilde{\Omega}_i^{\dagger}$ of the uncertainty set. For the first iteration, the original point $\widehat{\Omega}_i$ is the center point.

Given the center point $\widetilde{\Omega}_i^{\dagger}$, the concept of the sampling algorithm for a new row in the transition matrix starts with a random point Ω_i' on the simplex $\mathcal{H}$, where *i* denotes the *i*-th row of the transition matrix. If the random point Ω_i' is located in the uncertainty set Γ_i, then the new point $\widetilde{\Omega}_i(\mathcal{X})$ to test the inverse algorithm is the random point Ω_i'; otherwise, the random point Ω_i' serves as a starting point to search for the new testing point $\widetilde{\Omega}_i(\mathcal{X})$. When the random point Ω_i' is not in the uncertainty set Γ_i, the algorithm draws a line segment (a "diameter" of the uncertainty set) which connects the point Ω_i' and a point Ω_i'' on the "antipodal" boundary of $\mathcal{H}$ for which some element of Ω_i'' is equal to zero by passing through the center point $\widetilde{\Omega}_i^{\dagger}$. On this line segment, the sampling algorithm randomly selects a point $\widetilde{\Omega}_i$ and checks if the new sampled point $\widetilde{\Omega}_i$ is located within the uncertainty set Γ_i. If it is, then the coordinates of the point $\widetilde{\Omega}_i$ serve as the *i*-th row of the new transition matrix and serve as the new center point for the same row in the next iteration.

On the other hand, if the random point $\widetilde{\Omega}_i$ is not located in the uncertainty set Γ_i, then the sampling algorithm shrinks the "diameter" with endpoints Ω_i' and Ω_i'' by using $\widetilde{\Omega}_i$ to replace the diametric endpoint Ω_i' if $\widetilde{\Omega}_i$ lies on the line subsegment connecting Ω_i' to the center point $\Omega_i^{\dagger}$

of the uncertainty set; otherwise $\widetilde{\Omega}_i$ replaces the diametric endpoint Ω_i'' because $\widetilde{\Omega}_i$ must lie on the line subsegment connecting Ω_i'' to the center point $\Omega_i^\dagger$. Because the points $\widetilde{\Omega}_i$ and $\widetilde{\Omega}_i^\dagger$ are in between the two points Ω_i' and Ω_i'', we see that $\widetilde{\Omega}_i$ must be relatively closer to the center point $\widetilde{\Omega}_i^\dagger$ than one of the two points, Ω_i' or Ω_i''.

The algorithm picks a new random point $\widetilde{\Omega}_i$ between the updated Ω_i' and Ω_i'' and repeats the process of checking feasibility and movement until the point $\widetilde{\Omega}_i$ is located in the uncertainty set Γ_i. The movement of the point, Ω_i' or Ω_i'', is taken with respect to the center point $\widetilde{\Omega}_i^\dagger$; and the distance between Ω_i' and Ω_i'' is shortened after each movement so that Ω_i' and Ω_i'' progressively converge to "antipodal" boundary points of the uncertainty set Γ_i with respect to the "center point" $\Omega_i^\dagger$ of Γ_i until the point $\widetilde{\Omega}_i$ randomly sampled along the "diametric" line segment from Ω_i' to Ω_i'' finally satisfies $\widetilde{\Omega}_i \in \Gamma_i$ so that $\widetilde{\Omega}_i$ can be delivered as the next sampled point in Γ_i. The detail for searching a new row in a transition matrix is in Algorithm 5.2.1.

Algorithm 5.2.1 The uniform-spacings method for generating a new row in a transition matrix for $\mathcal{W}$ times

for $\mathcal{X} = 1$ to $\mathcal{W}$ step 1 **do**
  ▷ *Step 1* — Assign the center point:
  if $\mathcal{X} = 1$ **then**
    $\widetilde{\mathcal{Q}}_i^{\dagger} \longleftarrow \widehat{\mathcal{Q}}_i$;
  else
    $\widetilde{\mathcal{Q}}_i^{\dagger} \longleftarrow \widetilde{\mathcal{Q}}_i(\mathcal{X}-1)$.
  end if
  ▷ *Step 2* — Generate uniform spacings:
  $G \longleftarrow 0$.
  Generate ν random numbers, $\{U_i : i = 1, \ldots, \nu\} \overset{i.i.d.}{\sim}$ Uniform(0, 1).
  for $s = 1$ to ν **do**
    $F_s = -\ln(U_s)$,
    $G = G + F_s$.
  end for
  for $s = 1$ to ν **do**
    $F_s = \dfrac{F_s}{G}$.
  end for
  $\mathcal{Q}_i' = [F_1, F_2, \ldots, F_\nu]$.
  ▷ *Step 3* — Test for Feasibility:
  if $\mathcal{Q}_i' \in \Gamma_i$ **then**
    $\widetilde{\mathcal{Q}}_i(\mathcal{X}) = \mathcal{Q}_i'$;
  else
    Go to *Step 4*.
  end if
  ▷ *Step 4* — Searching $\mathcal{Q}_i''$ on the boundary:

for $j = 1$ to ν step 1 **do**

$\lambda_j \longleftarrow \dfrac{-\mathcal{Q}'_{ij}}{\widetilde{\mathcal{Q}}^{\dagger}_{ij} - \mathcal{Q}'_{ij}}$,

end for

$\lambda \longleftarrow \min\left\{\lambda_j : \lambda_j > 0 \;\&\; \mathcal{Q}'_i + \lambda_j\left(\widetilde{\mathcal{Q}}^{\dagger}_i - \mathcal{Q}'_i\right) \geqslant \vec{\mathbf{0}}_\nu\right\}$,

$\mathcal{Q}''_i \longleftarrow \mathcal{Q}'_i + \lambda\left(\widetilde{\mathcal{Q}}^{\dagger}_i - \mathcal{Q}'_i\right)$.

while $\widetilde{\mathcal{Q}}_i \notin \Gamma_i$ **do**

▷ *Step 5* — Compute the "location ratio" for $\widetilde{\mathcal{Q}}^{\dagger}_i$ along the line from $\mathcal{Q}'_i$ to $\mathcal{Q}''_i$:

$r^{\dagger} = \dfrac{\widetilde{\mathcal{Q}}^{\dagger}_{i,1} - \mathcal{Q}'_{i,1}}{\mathcal{Q}''_{i,1} - \mathcal{Q}'_{i,1}}$ (note $\mathcal{Q}''_{i,j} - \mathcal{Q}'_{i,j} \neq 0$ for $j \in \{1, \ldots, \nu\}$ when $\widetilde{\mathcal{Q}} \notin \Gamma_i$).

▷ *Step 6* — Sample $\widetilde{\mathcal{Q}}_i$ uniformly along the line which links $\mathcal{Q}'_i$ and $\mathcal{Q}''_i$:

$\widetilde{\mathcal{Q}}_i \longleftarrow \mathcal{Q}'_i + \widetilde{U}\left(\mathcal{Q}''_i - \mathcal{Q}'_i\right)$ where $\widetilde{U} \overset{i.i.d.}{\sim}$ Uniform(0, 1).

▷ Check feasibility:

if $\widetilde{\mathcal{Q}}_i \in \Gamma_i$ **then**

$\widetilde{\mathcal{Q}}_i(\mathcal{X}) \longleftarrow \widetilde{\mathcal{Q}}_i$;

else

Go to *Step 7*.

end if

▷ *Step 7* — Check the location of $\widetilde{\mathcal{Q}}_i$:

if $\widetilde{U} \leqslant r^{\dagger}$ **then**

$\mathcal{Q}'_i \longleftarrow \widetilde{\mathcal{Q}}_i$;

else

$\mathcal{Q}''_i \longleftarrow \widetilde{\mathcal{Q}}_i$.

end if

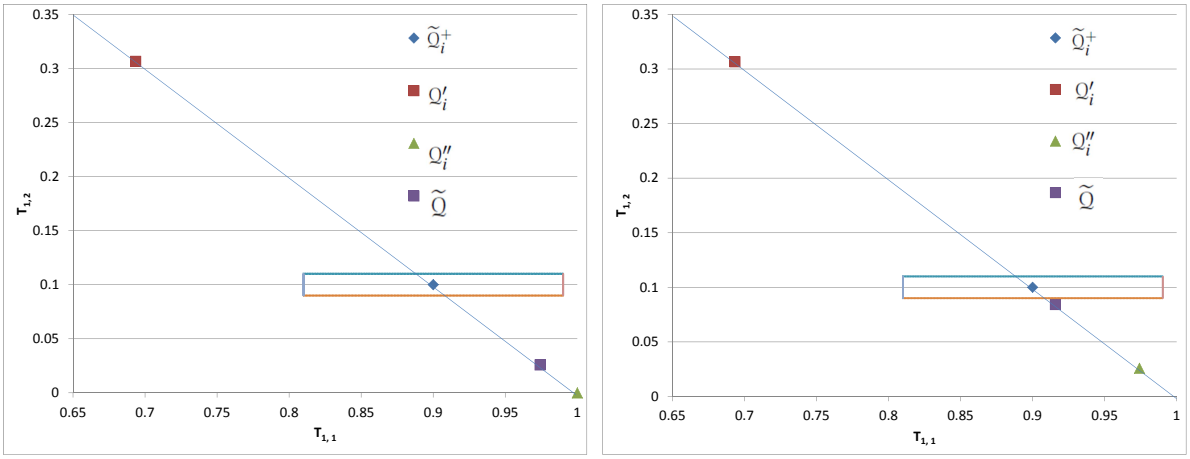
end while

end for

A two-dimensional example in Figure 5.8 illustrates the process of sampling one new row in the transition matrix. Assume $i = 1$ without loss of generality. The center point $\widetilde{\mathcal{Q}}_1^\dagger$ is $[0.9, 0.1]$ and the ranges of each element are $[0.81, 0.99]$ and $[0.09, 0.11]$ respectively, i.e., the uncertainty set Γ_1 is the part of the blue line segment inside the rectangle shown in Figure 5.8. In each of the graphs in Figure 5.8, the blue diamond represents the center point $\widetilde{\mathcal{Q}}_1^\dagger$; the red square represents the point $\mathcal{Q}_1'$; the green triangle represents the point $\mathcal{Q}_1''$; and the purple square represents the randomly selected point $\widetilde{\mathcal{Q}}_1$.

First, in Figure 5.8a, the algorithm selects a point $\mathcal{Q}_1' = [0.69343, 0.30657]$ on the simplex $\mathcal{H} := \{\pi \in \mathbb{R}^2 \mid \pi_1 + \pi_2 = 1, \pi_i \geqslant 0 \text{ for } i = 1, 2\}$. Since the point $\mathcal{Q}_1'$ is not located in the uncertainty set Γ_1, the sampling algorithm defines the point $\mathcal{Q}_1'' = [1, 0]$ on the boundary of the second dimension. Then, the sampling algorithm randomly selects a point $\widetilde{\mathcal{Q}}_1$ along the line segment that connects $\mathcal{Q}_1'$ and $\mathcal{Q}_1''$. From Figure 5.8a, we can see that $\widetilde{\mathcal{Q}}_1$ is located between the center point $\widetilde{\mathcal{Q}}_1^\dagger$ and the point $\mathcal{Q}_1''$. Hence, the sampling algorithm moves $\mathcal{Q}_1''$ to the location of $\widetilde{\mathcal{Q}}_1$ as shown in Figure 5.8b.

In the second round shown in Figure 5.8b, a new random point $\widetilde{\mathcal{Q}}_1$ between the two points $\mathcal{Q}_1'$ and $\mathcal{Q}_1''$ is selected. However, the point $\widetilde{\mathcal{Q}}_1$ is still not in the uncertainty set Γ_1. Then, the sampling algorithm has to move either $\mathcal{Q}_1'$ or $\mathcal{Q}_1''$. Since the point $\widetilde{\mathcal{Q}}_1$ is located between the center point $\widetilde{\mathcal{Q}}_1^\dagger$ and the point $\mathcal{Q}_1''$, $\mathcal{Q}_1''$ is moved again to the location of $\widetilde{\mathcal{Q}}_1$ shown in Figure 5.8c. In the third round shown in Figure 5.8c, the newly selected random point $\widetilde{\mathcal{Q}}_1$ is located in the uncertainty set Γ_1, i.e., the coordinates of the point $\widetilde{\mathcal{Q}}_1$ forms a new row of the testing transition matrix.



(a) First Round

(b) Second Round

(c) Third Round

Figure 5.8: A two dimensional example of the uniform-spacing method for sampling the uncertain set.

Results of the Sensitivity Analysis

Table 5.2 to Table 5.5 show the sensitivity analysis result of the imputed rewards $R_{i,\text{Ab}}^{\alpha(t)}$ for $i \in \{\text{P, C}\}$ for the age groups 40–44 and 75–79. Note that in Table 5.2–Table 5.5, the symbolism $\widetilde{\varphi} =$ Screening All represents the policy that screens at each decision epoch within the relevant age group. In each table, we compared the imputed reward $R^{\alpha}(t)_{i,\text{Ab}}$ for $i \in \{\text{P, C}\}$ for a specific age group $\alpha(t)$ with the sample results obtained by using the uniform spacings method to generate a random sample of size 50 from the uncertainty set centered on the imputed rewards $R_{i,\text{Ab}}^{\alpha(t)}$. In each table, Mean and H denote the sample mean and the half-length of a 95% CI estimator of $R^{\alpha}(t)_{i,\text{Ab}}$. It is important to observe that in each table the 95% CI includes the imputed value $R^{\alpha}(t)_{i,\text{Ab}}$, which at least partially validates the operation of the uniform spacings method. From the results, the imputed rewards of the do-nothing policy are all located on the same point of the reward space, $(R_{\text{P, Ab}}^{\alpha(t)}, R_{\text{C, Ab}}^{\alpha(t)}) = (1,0)$ because of the method which the inverse algorithm uses to address the degeneracy issue, i.e., the maximization of the summation of the differences between the value function of each alternative policy and the value function of the designated policy as formulated in Eq. (3.9).

For other policies, the results show the impact of a small change on the input data, the one-step transition probability matrix $T^{\alpha(t)}$ and the observation matrix $O^{\alpha(t)}$, on the imputed rewards. Given the random changes of at most $\pm 10\%$ in the input data, the impact on the imputed rewards of these two testing age groups are very small as measured by the relative deviation

$$D = \frac{\text{H}}{\text{Mean}} \tag{5.2}$$

of each 95% CI estimator of an imputed reward $R_{i,\text{Ab}}^{\alpha(t)}$. For example, from Table 5.2 we see that for policy φ_{ACS} the imputed value $R_{\text{P, Ab}}^{\alpha(t)}$=0.07514 for age group $\alpha(t) = 4$; and because D =0.00093, with 95% confidence we conclude that random variations of at most $\pm 10\%$ in

the input parameters of the inverse POMDP yield deviation of at most $\pm 0.09\%$ in the imputed value of $R^{\alpha(t)}_{\text{P, Ab}}$. Hence, we conclude that a small change in the input data does not change the impute rewards dramatically.

Table 5.2: Sensitivity analysis result of the imputed reward, $R^{\alpha(t)}_{\text{P, Ab}}$, for age group 40–44, $\alpha(t) = 4$.

Policy	$R^{\alpha(t)}_{\text{P, Ab}}$			
	φ_{NS}	$\widetilde{\varphi} = \varphi_{\text{ACS}}$	$\widetilde{\varphi} = \varphi_{\text{USP}}$	$\widetilde{\varphi}$ =Screening All
Original Data	1	0.07514	0.06798	0.04005
Mean	1	0.07539	0.06796	0.03995
Median	1	0.07539	0.06797	0.03996
Std. Dev.	0	0.00026	0.00005	0.00033
Minimum	1	0.07510	0.06785	0.03924
Maximum	1	0.07565	0.06807	0.04069
95% CI				
Half-length (H)	0	0.00007	0.00002	0.00009
D	0	0.00093	0.00029	0.00023

Table 5.3: Sensitivity analysis result of the imputed reward, $R^{\alpha(t)}_{\text{C, Ab}}$, for age group 40–44, $\alpha(t) = 4$.

Policy	$R^{\alpha(t)}_{\text{C, Ab}}$			
	φ_{NS}	$\widetilde{\varphi} = \varphi_{\text{ACS}}$	$\widetilde{\varphi} = \varphi_{\text{USP}}$	$\widetilde{\varphi}$ =Screening All
Original Data	0	0.97222	0.99	0.99
Mean	0	0.97427	0.99	0.99
Median	0	0.97333	0.99	0.99
Std. Dev.	0	0.00284	0	0
Minimum	0	0.97167	0.99	0.99
Maximum	0	0.98111	0.99	0.99
95% CI				
Half-length (H)	0	0.00081	0	0
D	0	0.00083	0	0

Table 5.4: Sensitivity analysis result of the imputed reward, $R^{\alpha(t)}_{\text{P, Ab}}$, for age group 75–79, $\alpha(t) = 11$.

Policy	$R^{\alpha(t)}_{\text{P, Ab}}$			
	φ_{NS}	$\widetilde{\varphi} = \varphi_{\text{ACS}}$	$\widetilde{\varphi} = \varphi_{\text{USP}}$	$\widetilde{\varphi}$ =Screening All
Original Data	1	0.06697	0.05973	0.03730
Mean	1	0.06694	0.05976	0.03733
Median	1	0.06697	0.05973	0.03720
Std. Dev.	0	0.00013	0.00034	0.00121
Minimum	1	0.06644	0.05928	0.03565
Maximum	1	0.06708	0.06034	0.03922
95% CI				
Half-length (H)	0	0.00004	0.00010	0.00034
D	0	0.00060	0.00167	0.00911

Table 5.5: Sensitivity analysis result of the imputed reward, $R^{\alpha(t)}_{\text{C, Ab}}$, for age group 75–79, $\alpha(t) = 11$.

Policy	$R^{\alpha(t)}_{\text{C, Ab}}$			
	φ_{NS}	$\widetilde{\varphi} = \varphi_{\text{ACS}}$	$\widetilde{\varphi} = \varphi_{\text{USP}}$	$\widetilde{\varphi}$ =Screening All
Original Data	0	0.98250	0.99100	0.99100
Mean	0	0.98210	0.99100	0.99100
Median	0	0.98225	0.99100	0.99100
Std. Dev.	0	0.00078	0.00000	0.00000
Minimum	0	0.98000	0.99100	0.99100
Maximum	0	0.98300	0.99100	0.99100
95% CI				
Half-length (H)	0	0.00022	0.00000	0.00000
D	0	0.00022	0	0

5.3 Conclusion

In this chapter, two parts of related research are presented: (a) the validation of the inverse algorithm; and (b) the sensitivity analysis of the imputed rewards. The purpose of the validation in Section 5.1 is to confirm that the imputed rewards which are obtained from the inverse algorithm do make the designated policy the optimal policy, while the purpose of the sensitivity analysis in Section 5.2 is to learn the effect of small changes in the input data on the imputed rewards (i.e., the one-step transition probability matrix and the observation probability matrix in this application).

In Section 5.1, the validation method is to solve a forward POMDP for breast cancer screening policies, which is described in Section 3.3. The policy evaluation method is selected to solve the forward POMDP problem so that the time-dependency feature of all policies can be maintained. In the validation analysis, the forward POMDP is solved in each of the following situations: (a) the strict equality constraint Eq. (2.45a) holds for the imputed rewards over the entire time horizon; and (b) the inequality constraint Eq. (2.46) with tolerance ε holds for the imputed rewards.

From the validation in situation (a) with the strict equality constraint Eq. (2.45a), the imputed rewards of the designated policy, do nothing, are located at the corner point of the reward space where the optimal policy $\breve{\varphi}$, which is obtained from the policy evaluation, is the do-nothing policy. The imputed rewards of all other designated policies includes some variability because of the tolerance. Hence, the validation in situation (b) uses the imputed rewards with some tolerance to search for the optimal policy. The second part of the validation shows the designated policy may not be the absolute optimal policy that is obtained from the policy evaluation method, but the differences between the value function for the absolute optimal policy $\breve{\varphi}$ and the value function for the designated policy $\widetilde{\varphi}$ are very small, i.e., approximately 10^{-6},

which may suggest that the policies are essentially equivalent to each other in terms of the value function.

In Section 5.2, we first introduce the sampling method for the one-step transition probability matrix and the observation matrix. Following the sampling method, the results of the sensitivity analysis are presented. The sensitivity analysis shows that changes of at most $\pm 10\%$ in the elements of the one-step transition probability matrix or the observation matrix lead to very small changes in the imputed rewards.

Chapter 6

Conclusions

Breast cancer is a fatal disease that can be prevented by the early detection and treatment. Different screening policies are recommended to detect breast cancer early. This dissertation aims to answer the following question: "What does a breast cancer screening policy say about a patient's health with regard to preventing breast cancer?" To address this problem, we develop an inverse POMDP algorithm for the time-dependent breast cancer screening policies. Using this inverse algorithm, we ananlyze of the advantages and disadvantages of time-dependent policies given the necessary optimality conditions of time-dependent policies.

In Chapter 2, we provide a detailed description of the inverse algorithm. A comprehensive discussion of a forward POMDP is necessary to build the foundation for the inverse algorithm. To accurately formulate the disease progression and the effect of the screening tests, we modify the sequence of events in a POMDP; and this complication leads to a revised formulation of the disease screening POMDP model. We then develop an inverse POMDP model for time-dependent policies.

The main idea of the inverse POMDP model is to maximize the summation of the differences in the value function solution for the designated policy and the value function solutions

for all other policies so that the value function for a designated policy is as far away from the value function for all alternative policies at the beginning of the time horizon. To maintain the optimality of the designated policy, constraints require that the value function for the designated policy should be better than or at least equal to the value function for all alternative policies at the beginning of the time horizon. The belief state space Π forms a part of the constraint set. This maximization problem is a nonlinear problem in the decision variables represented by the initial belief state $\pi^0 \in \Pi$ and the rewards $q^t_{i,\ell}(a)$ for $i \in \Omega$, $a \in \mathcal{A}$, $\ell \in \mathcal{Z}$, and $t \in \mathbb{T}$ because of the definition of the value function, Eq. (2.33). The structure of the belief state space Π allows us to covert the nonlinear maximization into several linear maximization problems by discretizing the belief state space and forming a finite subset (grid) $\mathscr{G}$ of initial belief states that are relevant to the population of US women between the ages of 25 and 100 who have not already been detected with breast cancer.

To resolve the degeneracy problem, we propose to search for the imputed rewards that maximize the summation of the differences of the value function results so that the value function for the designated policy is as far away as possible from the value function for all alternative policies. In addition, an equality constraint with tolerance ε is added to the rewards associated with a given observation, action, and state but with different decision epochs, so that the imputed rewards provide a more meaningful solution to the inverse breast cancer screening POMDP problem. A simple numerical example in Section 2.6 illustrates the inverse problem calculation step by step. This example shows that, without the equality constraint, the imputed rewards which are obtained from the inverse problem can be located in the extreme points of the reward space.

Chapter 3 describes the detailed formulation of both the forward POMDP and the inverse POMDP problem for time-dependent breast cancer screening policies. In addition to the inverse problem which is described in Section 2.5, several constraints are considered in this in-

verse POMDP: (a) the rewards of the forward POMDP should be between 0 and 1 because the definition of the rewards are the lifetime breast cancer mortality probabilities upon detecting the breast cancer in different stages, preclinical and clinical; and (b) the value function should be between 0 and 1 as well given the definition of the rewards.

Chapter 4 focuses on the numerical calculation and result of the inverse POMDP problem which is described in Chapter 3. To execute the inverse POMDP breast cancer screening model, we discuss the following issue: (a) the method for selecting the grid $\mathscr{G}$ of the relevant initial belief states; (b) the formulation for evaluating the imputed reward result, i.e., the expected number of mammograms; and (c) the data description of all data sources.

The numerical results shows the relationship between the imputed rewards associated with being detected in different states: the more screening tests a policy has, the higher the mortality probabilities are. However, the numerical results also suggest the benefit of one extra screening test does not provide as much benefit as the cost of the extra mammogram so that the existence of the trade-off between the imputed rewards and the expected number of mammograms is suggested. The timing of a screening test has a significant impact on the breast cancer mortalities associated with being detected in a particular cancer stage.

The operationally infeasible imputed rewards of some policies, such as the do-nothing policy, suggest a limitation of our inverse algorithm, which is that the imputed rewards are always located in the corner point of the feasible reward space because of the maximization of the summation of the differences between the value function for the designated policy and the value function for alternative policies. Different methods of resolving the degeneracy issue may need to be explored to address the operational infeasibility problem. A limitation which is suggested by the numerical results is the aggregating method of different initial belief states over the state space because more women are in a healthier belief state (such as $\pi^0 = [0.9, 0.1, 0, 0, 0]$) than the number of women in a belief state with high probabilities in state P and C (such as

$\pi^0 = [0.4, 0.2, 0.4]$). A weighted average method which considers the breast cancer risk factors may be a better method for aggregating the initial belief states.

The initial belief state with probability one in state NBC represents a special case because the infeasibility of some designated policies. The infeasibility in the LP problems comes from the constraint which ensures the optimality of the designated policy over the policy with a similar pattern of scheduled screening tests and one screening at the first decision epoch. Further study on this special case is necessary to understand the feasibility issue.

The tolerance ε relaxes the strict equality constraint on the imputed rewards so that the inverse problem for each combination of the designated policy and the initial belief state is feasible. The numerical results suggest the tolerance indicates that the younger the age group is, the larger the tolerance has to be so that the inverse problem is feasible for each combination of the designated policy and the initial belief state. Also, the tolerance varies across the designated policies that are chosen for analysis. To understand the behavior of the imputed rewards, more research is required, such as applying a different method to capture the relationship between different decision epochs.

Two main components, the validation and the sensitivity analysis, are discussed in Chapter 5. The purpose of the validation is to confirm that the imputed rewards, which are obtained from the inverse problem, do make the designated policy the optimal policy in the forward problem while the purpose of the sensitivity analysis is to study the impact on the imputed rewards when small changes are introduced in the input data.

The validation of the inverse problem is to solve the forward problem by using a policy evaluation method. The policy evaluation method ensures the time-dependency of the policies. We first perform the policy evaluation method under the assumption that the rewards associated with being detected in a particular state over the entire horizon are equal to each other. A total of 10201 different reward combinations over the entire reward space are used to solve the

forward problem. The results of this validation show that imputed reward combination which is obtained from the inverse problem for the do-nothing policy is located at the corner point in the region of the reward space where the optimal policy is the do-nothing policy within the region.

The second part of the validation is to use the imputed rewards with a tolerance in solving the forward problem. The results shows that the designated policy $\widetilde{\varphi}$ in the inverse problem may not coincide with the mathematically optimal policy $\breve{\varphi}$ in the forward problem, but the difference between the value function for the mathematically optimal policy $\breve{\varphi}$ and the value function for the designated policy $\widetilde{\varphi}$ is approximately 10^{-6}. The small difference suggests that the policies are essentially equivalent to each other in terms of the value function.

We perform a sensitivity analysis on the one-step transition probability matrix and the observation matrix with changes of at most $\pm 10\%$. A sampling method is introduced first to sample the new one-step transition probability matrix and the new observation matrix within the range. A total of 50 different combinations of the one-step transition probability matrix and the observation matrix are generated from the sampling method. The numerical results show the limited impact on the imputed rewards of the small changes in the one-step transition probability matrix and the observation matrix.

In summary, this dissertation provides a different perspective to evaluate the time-dependent breast cancer screening policies. The inverse algorithm provides a foundation of future research on obtaining the hard-to-measure information, such as the personal preferences of a patient for a specific policy, or the patient's concern about her total quality-adjusted life-years under different screening policies.

REFERENCES

[1] Ravindra K. Ahuja and James B. Orlin. Inverse optimization. *Operations Research*, 49(5):pp. 771–783, 2001.

[2] American Medical Association. Reports of The Council on Science and Public Health. In *Proceedings of the 2012 Annual Meeting of the House of Delegates*, 2012.

[3] M. Aoki. Optimal control of partially observable Markovian systems. *Journal of the Franklin Institute*, 280(5):367 – 386, 1965.

[4] A.R. Aro, S. Pilvikki Absetz, T.M. van Elderen, E. van der Ploeg, and L.J.Th. van der Kamp. False-positive findings in mammography screening induces short-term distress breast cancer-specific concern prevails longer. *European Journal of Cancer*, 36(9):1089 – 1097, 2000.

[5] K. Åström. Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174 – 205, 1965.

[6] Turgay Ayer, Oguzhan Alagoz, and Natasha K. Stout. OR Forum – A POMDP Approach to Personalize Mammography Screening Decisions. *Operations Research*, 60(5):1019–1034, 2012.

[7] D.P. Bertsekas. *Dynamic Programming and Stochastic Control*. Academic Press, 1976.

[8] BreastScreen Australia Program. BreastScreen Australia national policy. `http://www.cancerscreening.gov.au/internet/screening/publishing.nsf/Content/national-policy`, 2010.

[9] Robert Goodell Brown. *Smoothing, forecasting and prediction of discrete time series*. Englewood Cliffs, N. J., Prentice-Hall, 1963.

[10] D. Burton and Ph. L. Toint. On an instance of the inverse shortest paths problem. *Mathematical Programming*, 53:45–61, 1992.

[11] D. Burton and Ph. L. Toint. On the use of an inverse shortest paths algorithm for recovering linearly correlated costs. *Mathematical Programming*, 63:1–22, 1994.

[12] Carolina Mammography Registry. Available at: `http://www.unc.edu/cmr`. Accessed 30 June 2010.

[13] H. H. Chen, S. W. Duffy, and L. Tabár. A Markov chain method to estimate the tumour progression rate from preclinical to clinical phase, sensitivity and positive predictive value for mammography in breast cancer screening. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 45(3):pp. 307–317, 1996.

[14] Jaedeug Choi and Kee-Eung Kim. Inverse reinforcement learning in partially observable environments. *J. Mach. Learn. Res.*, 12:691–730, July 2011.

[15] Breast Cancer Surveillance Consortium. Screening performance. `http://breastscreening.cancer.gov/data/performance/screening/2009/perf_age.html`, National Cancer Institute, Bethesda, MD, 2009.

[16] D.M. Eddy. *Common Screening Tests*. American College of Physicians, Philadelphia PA, 1991.

[17] Zeynep Erkin, Matthew D. Bailey, Lisa M. Maillart, Andrew J. Schaefer, and Mark S. Roberts. Eliciting patients' revealed preferences: An inverse markov decision process approach. *Decision Analysis*, 7(4):358–365, 2010.

[18] Centers for Disease Control and National Center for Health Statistics Prevention. Underlying Cause of Death 1999-2009 on CDC Wide-ranging Online Data for Epidemiologic Research (WONDER) Online Database, released 2012. Accessed at http://wonder.cdc.gov/ucd-icd10.html on Aug 27, 2012 5:00:20 PM. Data for year 2009 are compiled from the Multiple Cause of Death File 2009, Series 20 No. 2O, 2012, Data for year 2008 are compiled from the Multiple Cause of Death File 2008, Series 20 No. 2N, 2011, data for year 2007 are compiled from Multiple Cause of Death File 2007, Series 20 No. 2M, 2010, data for years 2005-2006 data are compiled from Multiple Cause of Death File 2005-2006, Series 20, No. 2L, 2009, and data for years 1999-2004 are compiled from the Multiple Cause of Death File 1999-2004, Series 20, No. 2J, 2007.

[19] D. Goldfarb and A. Idnani. A numericallly stable dual method for solving strictly convex quadratic programs. *Mathematical Programming*, 27:1–33, 1983.

[20] Cheryl R. Herman, Harmindar K. Gill, John Eng, and Laurie L. Fajardo. Screening for preclinical disease: Test and disease characteristics. *American Journal of Roentgenology*, 179(4):825–831, 2002.

[21] C. Heuberger. Inverse combinatorial optimization: A survey on problems, methods, and results. *Journal of Combinatorial Optimization*, 8:329–361, 2004.

[22] Lindsay M. Howden and Julie A. Meyer. 2010 Census Briefs:Age and Sex Composition: 2010. Hyattsville, MD, 2011.

[23] N. Howlader, A.M. Noone, M. Krapcho, N. Neyman, R. Aminou, W. Waldron, S.F. Altekruse, C.L. Kosary, J. Ruhl, Z. Tatalovich, H. Cho, A. Mariotto, M.P. Eisner, D.R. Lewis, H.S. Chen, E.J. Feuer, and K.A. (eds) Cronin. *SEER Cancer Statistics Review, 1975-2009 (Vintage 2009 Populations)*. Bethesda, MD, 2011.

[24] National Cancer Institute. *What you need to know about breast cancer*. 2009. http://www.cancer.gov/cancertopics/wyntk/breast.

[25] National Cancer Institute. Surveillance epidemiology and end results (SEER). SEER stat fact sheets, cancer: Breast. http://seer.cancer.gov/statfacts/html/breast.html, 2012.

[26] Leslie Pack Kaelbling and Andrew W. Littman, Michael L. and Moore. Reinforcement learning: a survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

[27] A. W. Kuhn, H. W. and Tucker. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492, 1951.

[28] Carol H. Lee, D. David Dershaw, Daniel Kopans, Phil Evans, Barbara Monsees, Debra Monticciolo, R. James Brenner, Lawrence Bassett, Wendie Berg, Stephen Feig, Edward Hendrick, Ellen Mendelson, Carl D'Orsi, Edward Sickles, and Linda Warren Burhenne. Breast Cancer Screening with Imaging: Recommendations from the Society of Breast Imaging and the ACR on the Use of Mammography, Breast MRI, Breast Ultrasound, and Other Technologies for the Detection of Clinically Occult Breast Cancer. *Journal of the American College of Radiology : JACR*, 7:18–27, January 2010.

[29] Chertow Lee, C. P. and S. A. Zenios. A shadow price framework for quantifying health care demand, spending and disparity. *Revisions in Progress Management Science*, 2008.

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

[31] A.M. Miniño and S.L. Murphy. Death in the United States, 2010. In *NCHS data brief no. 99*. National Center for Health Statistics, Hyattsville, MD, 2012.

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

[33] National Cancer Insitute. Breast Cancer Risk Assessment Tool. Available at: http://www.cancer.gov/bcrisktool/Default.aspx.

[34] Andrew Y. Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *Proceeding of the 17th International Conference on Machine Learning*, pages 663–670. Morgan Kaufmann, 2000.

[35] NHS Breast Screening Programme. NHS Breast Screening Programme. http://www.cancerscreening.nhs.uk/breastscreen/screening-programme.html, 2011.

[36] American College Of Obstetricians and Gynecologists. Practice bulletin no. 122: Breast cancer screening. *Obstetrics & Gynecology*, 118(2):372–382, August 2011.

[37] Surveillance Research Program. *An interactive tool for access to SEER cancer statistics*. National Cancer Institute. `http://seer.cancer.gov/faststats`. Accessed on 8-27-2012.

[38] M.L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. Wiley series in probability and statistics. Wiley-Interscience, 2005.

[39] Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *20th International Joint Conference Artificial Intelligence*, pages 2587–2591, 2007.

[40] Detlef Rhenius. Incomplete information in Markovian decision models. *The Annals of Statistics*, 2(6):1327–1334, 1974.

[41] Stuart Russell. Learning agents for uncertain environments (extended abstract). In *Proceedings of the eleventh annual conference on Computational learning theory*, COLT' 98, pages 101–103, New York, NY, USA, 1998. ACM.

[42] R. D. Smallwood and E. J. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973.

[43] American Cancer Society. *Breast Cancer Facts and Figures 2011–2012*. Atlanta, GA, 2012. `http://www.cancer.org/acs/groups/content/@epidemiologysurveilance/documents/document/acspc-030975.pdf`.

[44] American Cancer Society. *Cancer Facts and Figures 2013–2014*. Atlanta, GA, 2013. `http://www.cancer.org/acs/groups/content/@research/documents/document/acspc-040951.pdf`.

[45] E. J. Sondik. *The optimal control of partially observable Markov processes*. PhD thesis, Stanford Universtiy, 1971.

[46] László Tabár, Bedrich Vitak, Hsiu-Hsi Chen, Stephen W. Duffy, Ming-Fang Yen, Ching-Feng Chiang, Ulla Brith Krusemo, Tibor Tot, and Robert A. Smith. The Swedish two-county trail twenty years later: Updated mortality results and new insights from long-term follow-up. *Radiologic Clinics of North America*, 38(4):625 – 651, 2000.

[47] A. Tarantola. *Inverse problem theory : methods for data fitting and model parameter estimation*. Elsevier, Amsterdam, 1987.

[48] A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics, 2005.

[49] The Canadian Task Force on Preventive Health Care. Recommendations on screening for breast cancer in average-risk women aged 40-74 years. *Canadian Medical Association Journal*, 183:19912001, November 2011.

[50] U.S. Preventive Services Task Force. Screening for Breast Cancer, Topic Page. http://www.uspreventiveservicestaskforce.org/uspstf/uspsbrca.htm, USPSTF Program Office 540 Gaither Road, Rockville, MD 20850, July 2010.

[51] Jenny Wu, Matti Hakama, Ahti Anttila, Amy Yen, Nea Malila, Tytti Sarkeala, Anssi Auvinen, Sherry Chiu, and Hsiu-Hsi Chen. Estimation of natural history parameters of breast cancer based on non-randomized organized screening data: subsidiary analysis of effects of inter-screening interval, sensitivity, and attendance rate on reduction of advanced cancer. *Breast Cancer Research and Treatment*, 122:553–566, 2010.

[52] Bonnie C. Yankaskas, Sebastien Haneuse, Julie M. Kapp, Karla Kerlikowske, Berta Geller, Diana S. M. Buist, and for the Breast Cancer Surveillance Consortium. Performance of first mammography examination in women younger than 40 years. *Journal of the National Cancer Institute*, 102(10):692–701, 2010.

[53] Jianzhong Zhang and Zhenhong Liu. Calculating some inverse linear programming problems. *Journal of Computational and Applied Mathematics*, 72(2):261 – 273, 1996.

[54] Jianzhong Zhang and Zhenhong Liu. A further study on inverse linear programming problems. *Journal of Computational and Applied Mathematics*, 106(2):345 – 359, 1999.

[55] Shengfan Zhang and Julie S. Ivy. Analytic Modeling of Breast Cancer Spontaneous Regression. In *Proceedings of the 2012 Industrial and Systems Engineering Research Conference*, Orlando, FL, 2012.

[56] Shengfan Zhang, Julie S. Ivy, Kathleen M. Diehl, and Bonnie C. Yankaskas. The association of breast density with breast cancer mortality in african american and white women screened in community practice. *Breast Cancer Research and Treatment*, 137(1):273–283, 2013.

[57] Shao Zhifei and Er Meng Joo. A review of inverse reinforcement learning theory and recent advances. In *Evolutionary Computation (CEC), 2012 IEEE Congress on*, pages 1–8, June 2012.