Stochastic Models for Improving Screening and Surveillance Decisions for Prostate Cancer Care

by

Christine Barnett

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in The University of Michigan
2017

Doctoral Committee:

Professor Brian T. Denton, Chair
Associate Professor Mariel S. Lavieri
Professor Lawrence M. Seiford
Assistant Professor Scott A. Tomlins

Christine L. Barnett

clbarnet@umich.edu

ORCID iD: 0000-0002-1465-7623

DEDICATION

For Jon, Amy, and Kate.

ACKNOWLEDGEMENTS

First, I would like to thank my advisor, Dr. Brian Denton, for his guidance and support over the past five years. I am grateful to have had him as a mentor through this experience, and feel prepared for the next stage in my career thanks to his encouragement and guidance. This work was supported in part by the National Science Foundation through Grant Number CMMI 0844511 and by the National Science Foundation Graduate Research Fellowship under Grant Number DGE 1256260.

I would like to thank my committee members Dr. Mariel Lavieri, Dr. Lawrence Seiford, and Dr. Scott Tomlins for serving on my committee and providing me with career advice and helpful feedback on my research. In addition, I would like to thank our collaborators from Michigan Medicine, Dr. Gregory Auffenberg, Dr. Matthew Davenport, Dr. Jeffrey Montgomery, Dr. James Montie, Dr. Todd Morgan, Dr. Scott Tomlins, and Dr. John Wei for their invaluable clinical perspective and for teaching me about the intricacies of prostate cancer screening and treatment decisions.

Finally, I would like to thank my friends and family for their support throughout graduate school. Thank you to my mom, Amy, for her unwavering love and support throughout my life. I owe any of my successes to her. Thank you to my sister, Kate, for believing in me. Thank you to my husband, Jonathan, for his everlasting support and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF ABBREVIATIONS

AUA American Urological Association

CT computed tomography

DRE digital rectal examinations

ERSPC European Randomized Study of Screening for Prostate Cancer

hK2 human kallikrein-2

ICER incremental cost-effectiveness ratio

MiPS Mi-Prostate Score

MISCAN Microsimulation Screening Analysis

MRI magnetic resonance imaging

PCA3 prostate cancer antigen 3

phi prostate health index

PI-RADS prostate imaging reporting and data system

POMDP partially observable Markov decision process

PSA prostate-specific antigen

QALY quality-adjusted life year

SEER Surveillance, Epidemiology, and End Results

T2:ERG TMPRSS2:ERG

ABSTRACT

Stochastic Models for Improving Screening and Surveillance Decisions for Prostate Cancer Care

by

Christine Barnett

Chair: Brian T. Denton

Recent advances in the development of new technologies for the early detection and treatment of cancer have the potential to improve patient survival and lower the cost of treatment by catching cancer at an early stage. However, there is little research investigating the health and economic implications of these new technologies. For example, magnetic resonance imaging (MRI) and new biomarker tests have been proposed as potential minimally invasive ways to achieve early detection of prostate cancer. These new technologies vary in their sensitivity and specificity leading to both false-positive and false-negative results that can have serious health implications for patients. Moreover, due to the high cost and imperfect nature of these new tests, whether and when to use these tests is unclear.

We present stochastic models for prostate cancer disease onset and progression that incorporates partial observability of a patient's prostate cancer health status. We used statistical learning algorithms and clinical datasets combined with expert clinical knowledge of urologists at the University of Michigan to estimate and validate the models. The models can simulate progression through prostate cancer states to

mortality from prostate cancer or other causes for a population of patients. New technologies, such as MRI and biomarker tests, are incorporated into the model using a probabilistic representation of test outcomes to represent the information these tests provide about the true health status of the patient. Since these technologies can be used in varying ways, the choice of tests and optimal times to initiate tests are treated as decision variables in the model. We calibrated and validated our models using several data sources and subsequently used our models to design optimal testing strategies that trade-off the harms and benefits of using these new technologies.

Our results show that these new technologies can lead to significantly improved health outcomes and they are cost-effective relative to established norms for societal willingness-to-pay. We have also used these models to provide important insights about the optimal timing of prostate biopsies for men with low-risk prostate cancer undergoing active surveillance. By using new technologies to better select men for biopsy and by improving active surveillance strategies, physicians can reduce the harms of prostate cancer screening (e.g., unnecessary biopsies and overtreatment of low-risk disease) while continuing to reduce prostate cancer deaths through screening and early detection. The methodological approaches we present in this thesis could be applied to many other chronic diseases, including bladder, breast, and colorectal cancer.

CHAPTER I

Introduction

Cancer screening has the potential to improve patient survival and lower the cost of treatment by detecting cancer at an early stage when health outcomes are most favorable for patients. However, there are several challenges associated with screening for cancer. For example, the tests used for cancer screening are imperfect and there are harms associated with the screening process. Additionally, multiple grades of cancer indicate that, while cancer screening could save the life of a patient with high-grade cancer, it is unlikely to benefit patients with low-grade cancers. Moreover, there is uncertainty about progression of cancers over time and uncertainty about the benefits and side effects of treatment. Finally, the benefits of cancer screening depend on all-cause mortality, since patients with a lower expected lifespan are unlikely to receive the benefits of screening. Due to these issues, decisions about cancer screening are challenging.

Prostate cancer is the ideal context to explore these challenging problems because of (1) its societal importance (one in six men are diagnosed in the United States); (2) the prostate-specific antigen (PSA) test is an existing biomarker that is in common use; and (3) many new prostate cancer screening biomarkers have recently been developed. Additionally, prostate cancer can have slow progression, and patients with different grades of prostate cancer have significantly different treatment options and

survival outcomes.

Prostate cancer is the most common cancer among men in the United States. The risk of developing prostate cancer varies among patients depending on many factors. For example, patients who are African-American or have a family history of prostate cancer are considered to be at a higher risk for the disease. Since prostate cancer is asymptomatic at early stages, some physicians screen their patients for prostate cancer using digital rectal examinations (DRE) and the PSA test. If the results of these tests are "suspicious", a biopsy is performed.

Two recent clinical trials to evaluate the effectiveness of PSA screening for preventing prostate cancer death have resulted in contradictory findings. The European Randomized Study of Screening for Prostate Cancer (ERSPC), *Schröder et al.* (2009, 2012, 2014), randomized 162,387 men to either a screening group or a control group at seven centers in European countries. The relative risk after 11 years of follow-up was a statistically significant 0.79 between the screening and control arms, interpreted as a 20% risk reduction in prostate cancer mortality due to early diagnosis and treatment. The Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial, *Andriole et al.* (2009), randomized 76,693 men to either a screening group or a control group at 10 centers in the United States. The relative risk of prostate cancer mortality after 13 years of follow-up was a non-statistically significant 1.09 between the screening and control arms, showing no benefit from early diagnosis and treatment (*Andriole et al.* (2012)). These conflicting findings suggest that randomized control trials are not the ideal way to evaluate screening policies. Due to these conflicting findings, there is disagreement about if and when men should be screened. For example, the American Urological Association (AUA) recommends PSA screening for men from ages 55 to 69 with two year intervals (*Carter et al.* (2013)), while the U.S. Preventive Services Task Force recommends against PSA screening due to the *over-treatment* and *unnecessary biopsies* that have been attributed to wide-spread PSA screening (*Moyer* (2012)).

Thus, in this thesis we study optimal approaches to conduct screening in a manner that reduces unnecessary biopsies and the overtreatment of low-grade prostate cancer.

Recent advances in the development of new technologies for the early detection of prostate cancer have the potential to supplement the PSA test and improve patient survival by catching cancer at an early stage when health outcomes are most favorable for patients. However, there is little research investigating the long-term health and economic implications of these new technologies. For example, magnetic resonance imaging (MRI) and new biomarker tests have been proposed as potential minimally invasive ways to achieve early detection of prostate cancer, but whether and when to use them is unclear due to the high cost and imperfect nature of these tests.

Several new diagnostic prostate cancer biomarkers have been recently discovered (*Makarov et al.* (2009); *Tosoian et al.* (2016)). Some of these biomarkers are PSA derivatives, such as free PSA and [-2]proPSA. Some of the biomarkers are based on combinations of serum markers, such as the prostate health index (phi), which uses a combination of total PSA, free PSA, and [-2]proPSA to generate a score (*Bryant and Lilja* (2014); *Catalona et al.* (2011)), and the 4Kscore, which uses a panel of total PSA, free PSA, intact PSA, and human kallikrein-2 (hK2) to estimate a patient's risk of high-grade cancer (Gleason score ≥ 7) on biopsy. Other molecular biomarkers include prostate cancer antigen 3 (PCA3) and TMPRSS2:ERG (T2:ERG), which are detectable in post-DRE urine (*Bussemakers et al.* (1999); *Salagierski and Schalken* (2012); *Truong et al.* (2013); *Tomlins et al.* (2005); *Brenner et al.* (2013); *Pettersson et al.* (2012); *Young et al.* (2012)). The Mi-Prostate Score (MiPS) early detection test combines a patient's serum PSA, urine PCA3 score, and urine T2:ERG score into a single multivariate regression model to estimate individualized risk estimates for all prostate cancer and high-grade prostate cancer (*Tomlins et al.* (2016)). These tests vary in the outcome they predict (all-cancer versus high-grade cancer) and in their sensitivities and specificities. No previous study has compared these biomarkers

to determine which characteristics achieve optimal long-term health outcomes in the context of early detection of prostate cancer.

MRI has recently been proposed as another potential minimally invasive way to achieve early detection of prostate cancer. MRI has higher sensitivity and specificity to high-grade disease than molecular biomarkers, so MRI could potentially reduce overtreatment by preferentially detecting intermediate- and high-grade cancers (*Siddiqui et al.* (2015); *Meng et al.* (2016); *Oberlin et al.* (2016); *Siddiqui et al.* (2016)). However, MRI is more costly than molecular biomarkers and there is limited evidence for its effectiveness as an intermediate test in patients being screened for prostate cancer. Moreover, there are multiple ways to use MRI in a screening setting, and it is not clear which is best.

Based on biomarker test results, MRI results, or other clinical findings, a patient may be recommended to have a biopsy. During a biopsy, a hollow core needle is used to remove between 6 and 24 (usually 12) core samples of tissue from the prostate to determine if the tissue is malignant. Biopsies have a specificity close to 1 and a sensitivity of approximately 0.8 according to studies in the literature (*Terris* (1999)). Biopsies are painful, may cause bleeding and infection, and can be a source of anxiety for patients (*Wade et al.* (2013)). If cancer cells are found upon evaluation of the biopsy by a pathologist, the cells are given a Gleason score. The two most common tissue patterns of the prostate tissue (obtained during the biopsy) receive a grade between 1 and 5. This grade rates how different the cancer cells are from normal cells. These two grades are added together to obtain a Gleason score between 2 and 10. A higher Gleason score indicates that the tumor is more likely to grow and spread quickly.

Once a patient has been diagnosed with prostate cancer, a physician needs to determine the stage of the disease. During clinical staging, it is important to detect metastatic disease (i.e. when the cancer has spread to other parts of the body),

because this determines treatment options. Thus, a computed tomography (CT) scan or bone scan may be performed to screen patients with a high risk of metastatic disease. CT scans and bone scans are used to detect whether the cancer has spread to the lymph nodes and the bones, respectively.

Following diagnosis there are many factors that influence treatment decisions, including the stage and grade of the cancer, the age and life expectancy of the patient, possible side effects, and whether the patient has other health conditions. If a patient is diagnosed with metastatic cancer, they will most likely receive hormone therapy or chemotherapy. For localized cancer there are several treatment options.

Radical prostatectomy is the surgical removal of the prostate gland and surrounding tissue, and is an appropriate treatment option when the cancer is contained within the prostate. This procedure can be done as an open surgery or laparoscopically, which affects the recovery time. Radical prostatectomy has the same risks as other major surgeries including infection, blood loss, and heart problems. Some possible long-term side effects include impotence, urinary incontinence, and damage to the urethra or rectum; however, it is usually effective in curing early-stage prostate cancer.

For men with low-grade cancer, there are currently two types of observational treatment strategies, known as *expectant management*, that can serve as alternatives to aggressive immediate treatments. *Watchful waiting* is an expectant management option that delays curative treatment until symptoms arise. *Active surveillance* is an expectant management option for patients with low-risk prostate cancer that delays and possibly avoids curative treatment until there is evidence that the disease has progressed; however, the patient must undergo repeated biopsies and there is still the potential for the cancer to progress due to the imperfect nature of surveillance biopsies. Expectant management has the benefit that it allows men to delay and possibly avoid the side effects of curative treatment; however, the optimal timing of

biopsies during active surveillance is unknown.

The two main harms of prostate cancer screening are unnecessary biopsies caused by false-positive PSA results and the overtreatment of low-risk prostate cancer with harsh curative treatments. The goal of this thesis is to discover ways to reduce the harms of prostate cancer screening by providing insights on how new technologies can be used with the established PSA test to better select men for prostate biopsy, as well as improve the understanding of risks associated with active surveillance strategies to encourage patients and physicians to reduce overtreatment in favor of active surveillance. By reducing the harms associated with screening, physicians can safely screen men for prostate cancer and reduce prostate cancer deaths. To achieve these goals we describe stochastic models we developed and validated that cover the complete life cycle of prostate cancer from early ages when the probability of prostate cancer is low, through potential onset and progression of cancer, and subsequent treatment. These models are used to answer key questions about prostate cancer, such as whether, when, and how to use new technologies or procedures judiciously to improve quality of life and lifespan for men.

This dissertation is structured as follows. In Chapter II we present a literature review on simulation and optimization of cancer modeling. In Chapter III, we present a prostate cancer simulation model that we used to evaluate a wide range of new biomarkers for the early detection of prostate cancer in patients with elevated PSA. In Chapter IV, we present cost-effectiveness analysis of using MRI for the early detection of prostate cancer in men with elevated PSA. In Chapter V, we develop a partially observable Markov decision process (POMDP) to determine optimal prostate biopsy decisions using new biomarkers. In Chapter VI, we use longitudinal data from the Johns Hopkins Active Surveillance study to create a hidden Markov model that we used to develop optimal biopsy follow-up schedules for patients with low-risk prostate cancer. Finally, in Chapter VII we discuss the main findings from this dissertation

and discuss potential future research extensions.

CHAPTER II

A Literature Review

The scope of this literature review includes simulation and optimization models for cancer screening. It is a narrative review that describes related literature. We provide a more specific discussion of the most relevant literature in each of the following chapters.

2.1 Simulation Models

Many models have studied the lead time and overdetection of prostate cancer, where the lead time of prostate cancer refers to the advanced time of diagnosis resulting from the use of biomarker tests, and overdetection of prostate cancer refers to patients who are diagnosed with prostate cancer that would not have been diagnosed in the absence of screening. *Etzioni et al.* (2002) developed a computer simulation model of PSA testing and prostate cancer to provide estimates of overdiagnosis due to PSA screening. They found that the majority of screen-detected cancers between 1988 and 1998 would have presented clinically in the patient's lifetime, and therefore were not overdiagnosed. *Draisma et al.* (2003) developed simulation models using the outcomes of the Rotterdam section of the ERSPC to predict mean lead times and overdetection rates of different screening policies. Based on their results, they concluded that a screening interval of more than one year would be optimal. *Tsodikov*

et al. (2006) used a statistical model to estimate prostate cancer screening characteristics, such as lead time and overdiagnosis, to try to find a connection between the onset of PSA screening and population responses observed in Surveillance, Epidemiology, and End Results (SEER) registry data. *Draisma et al.* (2009) presented three independent prostate cancer models that were developed using SEER registry data to estimate the lead time of the disease, as well as the overdetection rate. They found that their estimated lead times were similar, but differed based on the definition used, concluding that the definition of lead time in models should always be clearly defined. *Savage et al.* (2010) developed an empirical lead time distribution between an elevated PSA ($\geq$ 3 ng/mL) and subsequent prostate cancer diagnosis, and found that there was wide variation in lead times, with longer lead times having a lower risk of high-grade disease. *Gulati et al.* (2014) developed a nomogram that provides patients with individualized estimates of the risk that their screen-detected prostate cancer is overdiagnosed, based on the patient and tumor information known at diagnosis. Important factors in determining the chance of overdiagnosis are age, Gleason score, and PSA at diagnosis.

There have also been many models that have estimated the effect of widespread PSA-screening on prostate cancer statistics. *Etzioni et al.* (2008a) developed a fixed-cohort simulation model of prostate cancer screening to determine the impact of PSA screening on the incidence of advanced stage prostate cancer in the United States. They determined that PSA screening accounted for 80% of the observed decline in distant stage incidence, but concluded that other factors have most likely also contributed to the decline, such as improved treatment modalities and increased awareness. *Etzioni et al.* (2008b) presented two mathematical models that use SEER registry data to project mortality increases in the absence of screening and decreases in the presence of PSA screening. The models found that 45% to 70% of the observed decline in prostate cancer mortality in the United States could be attributed to PSA

screening.

There have also been studies that compared the effectiveness of alternative PSA screening policies (e.g. annual screening or biennial screening, varying starting age for screening). *Ross et al.* (2000) compared prostate cancer mortality, PSA testing rates, and biopsy rates for several different PSA screening strategies using a Markov Chain-based Monte Carlo simulation. They found that a policy that begins earlier than age 50 and screens biennially instead of annually would perform better than screening annually beginning at age 50. *Gulati et al.* (2013) developed a microsimulation model of prostate cancer to evaluate the effectiveness of PSA-based prostate cancer screening strategies. They evaluated strategies recommended by guidelines and 32 combinations of two ages to start screening, two ages to end screening, two screening intervals, and four PSA thresholds for biopsy. They concluded that PSA screening strategies that have higher thresholds for biopsy referral for older men and that screen men with low PSA levels less frequently performed better than standard screening in terms of minimizing both the harms of screening and prostate cancer deaths.

Gulati et al. (2011) developed a simulation model to project long-term estimates of the number needed to screen and the number needed to treat to prevent one prostate cancer death with PSA screening from ERSPC, and found that their long-term estimates are much more favorable than the previous short-term estimates published by *Schröder et al.* (2009). *Heijnsdijk et al.* (2012) presented a Microsimulation Screening Analysis (MISCAN) model based on ERSPC follow-up data to predict various long-term prostate cancer screening outcomes. *Heijnsdijk et al.* (2015) used the MISCAN model to evaluate the cost-effectiveness of prostate cancer screening, and concluded that prostate cancer screening can be cost-effective when the patient receives two or three screenings between ages 55 and 59 years. *Roth et al.* (2016) also studied the cost-effectiveness of PSA screening using a microsimulation model; however, their cohort was based on a US population. *Roth et al.* (2016) concluded that for PSA

screening to be cost effective, it needs to be used conservatively with conservative management approaches for low-risk disease.

Underwood et al. (2012) developed a genetic algorithm simulation optimization model based on a non-stationary finite horizon Markov chain to develop a PSA screening policy that maximizes expected quality-adjusted life years (QALYs). The genetic algorithm uses tournament selection to choose parents, two-point crossover to create offspring, and elitism and mutation to create the next generation. The policy generated by the genetic algorithm performed better than previously published policies in terms of QALYs. The simulation optimization model indicated that patients should be screened more aggressively for a shorter period of time.

Several studies have developed models to evaluate new technologies for the early detection of prostate cancer. *Heijnsdijk et al.* (2016) used MISCAN to evaluate the effects of the new biomarker, phi, on prostate cancer screening. They concluded that the use of phi in patients with elevated PSA substantially reduced the number of negative biopsies and improved the cost-effectiveness of prostate cancer screening. *Birnbaum et al.* (2015) used a simulation model to evaluate the effect of the new biomarker, PCA3, on prostate cancer screening, and found that supplementing PSA with the PCA3 test significantly reduced adverse screening outcomes. *Willis et al.* (2014) performed a clinical decision analysis and *de Rooij et al.* (2014) performed a cost-effectiveness analysis of using MRI followed by targeted prostate biopsy for early detection of prostate cancer. *Willis et al.* (2014) found that using MRI resulted in fewer biopsies and more clinically significant cancer diagnoses, while *de Rooij et al.* (2014) found that using MRI resulted in similar costs to the current standard of care while achieving more QALYs.

Active surveillance is a treatment option for low-risk prostate cancer patients that delays or avoids curative treatment until there is evidence of disease progression; however, the patient must receive serial prostate biopsies, there is the potential

for misclassification at diagnosis due to undersampling at biopsy, and the optimal timing of biopsies during surveillance is unknown. *Inoue et al.* (2014) used serial prostate biopsy data from the Johns Hopkins Active Surveillance study to explore whether Gleason score upgrading during active surveillance was due to misclassification or true grade progression. *Inoue et al.* (2014) developed a statistical model that estimated true grade progression rates, while accounting for misclassification due to undersampling at diagnosis biopsy. They applied their model to serial prostate biopsy results for patients enrolled in active surveillance at Johns Hopkins, and concluded that tumor grade can progress in low-risk prostate cancer patients. In a related bladder cancer study, *Zhang et al.* (2013) studied the optimal timing of cystoscopies for patients with low-grade noninvasive bladder cancer. They developed a partially observable Markov model to estimate QALYs, expected lifelong progression probability, and lifetime number of cystoscopies for varying surveillance strategies.

There have also been a number of related studies in the context of breast cancer. *Maillart et al.* (2008) considered the problem of how often premenopausal and post-menopausal women should receive mammography screening for breast cancer. They formulated a partially observable Markov chain for breast cancer progression including five states. The model assumes that patients are only diagnosed with breast cancer during routine screening. Thus, the model provides a conservative, worst-case scenario analysis, since some patients would develop symptoms and be diagnosed between screening. Unlike previous models, their model addressed multiple age-based dynamics of breast cancer screening. Since adherence to the current routine policy recommendation is low, *Maillart et al.* (2008) only considered "two-phase" policies, which consist of only two different screening intervals. The experimental design of *Maillart et al.* (2008) resulted in 1223 enumerated possible policies, and sample-path enumeration was used to generate a frontier of efficient policies by balancing the lifetime mortality risk and the expected number of mammograms throughout a patient's

lifetime. The current recommendation of annual mammograms beginning at age 40 is dominated by less than 1% by the frontier efficient policies. *Maillart et al.* (2008) provides a range of efficient and less than 1% efficient policies that patients can select from based on their personal preferences. It was determined that screening should start relatively early and end relatively late in life regardless of the interval between mammograms.

Tejada et al. (2015) developed a natural history simulation model of breast cancer in a simulated population of women in the United States over age 65. *Tejada et al.* (2014) used the natural histories of the simulated population to model breast cancer screening for women over age 65. They combined discrete-event simulation and system dynamics submodels to compare screening policies based on overall cost-effectiveness, cost incurred, and the numbers of life-years and QALYs saved. They considered interval screening policies (i.e. one policy for the entire population), risk-based screening policies, and factor-based screening policies accounting for factors such as age, race and body mass index. Their final recommendation is annual screening between ages 65 and 80.

Mandelblatt et al. (2016) was a collaboration of six simulation models that evaluated breast cancer screening outcomes, accounting for recent advances in mammography and treatment. They found that biennial breast cancer screening with mammography is efficient for average-risk patients, and that optimal starting ages and screening intervals depend on patient characteristics and preferences. *Ayer et al.* (2015) found that since adherence to mammography screening policies is imperfect and heterogeneous, an aggressive screening strategy recommending annual screening to the general population should be recommended.

Similar studies have been published related to lung cancer screening. For example, *de Koning et al.* (2014) conducted a comparative modeling study using five independent models, and suggested annual lung cancer screening with CT for patients age

55 to 80 with ≥ 30 pack-years of smoking. This screening policy resulted in 50% of cases diagnosed at an early stage and a 14% reduction in lung cancer mortality.

2.2 Optimization Models

There have been a number of previous studies that have used optimization models to investigate screening and treatment decisions for cancer in the context of imperfect screening tests.

Zhang et al. (2012a) developed a nonstationary POMDP for prostate biopsy referral decisions that maximizes expected QALYs, and found that the decision of when to stop screening is highly dependent on the patient and the disutility of life after treatment. It is proven that there exists a control-limit type policy, and the computational experiments performed indicate that there is a nondecreasing belief threshold in age. Sufficient conditions for discontinuing PSA screening for older patients are presented. *Zhang et al.* (2012b) expanded on this work by including PSA screening decisions about whether and when to screen over the course of a patient's lifetime. *Zhang* (2011) presents a POMDP for prostate cancer treatment decisions, incorporating active surveillance. At each decision epoch, the patient can wait, receive a PSA test or biopsy, or receive definitive treatment. The underlying partially observable Markov chain formed the initial basis of the model that we developed and validated, which we describe in Chapter III.

Lavieri et al. (2012) investigated a model to optimize the timing of when to begin radiation therapy in prostate cancer patients based on the patient's PSA level. Their model balances the risk of beginning radiation therapy too soon before hormone therapy has achieved its maximum effect with the risk of beginning radiation therapy too late when tumor cells become resistant to treatment.

Simmons Ivy et al. (2009) developed a simulation model that combines statistical control and a POMDP to quantify the impact of variability and noise on patient

outcomes in breast cancer decision making. They found that variability among radiologists in interpreting mammography results has the largest impact on a patient's outcomes. Thus, reducing this variability should be a primary goal to improve women's healthcare.

Chhatwal et al. (2010) developed a finite-horizon discrete-time Markov decision process for breast biopsy referral decisions. The model takes into account a patient's mammographic features and demographic factors. The optimal policy shows that there is a control-limit type policy based on the patient's breast cancer risk and nondecreasing control-limits with age. Using clinical data, their model outperforms radiologists in the biopsy decision-making problem, which may have the ability to reduce the variability noted in *Simmons Ivy et al.* (2009).

Ayer et al. (2012) was the first paper to consider personalized mammography-screening policies, based on both static and dynamic risk factors. They develop a discrete-time, finite-horizon POMDP model to determine the optimal policy for an individual patient in terms of maximizing the total expected QALYs. The POMDP consists of six states, including multiple cancer states. Every six months, based on the woman's current risk of breast cancer a decision is made about whether the patient should receive a mammogram or should wait six months. In this model, if a mammogram comes back positive, the patient receives a biopsy, which is assumed to be perfect. *Ayer et al.* (2012) also incorporates the possibility of self-detection into the problem, and shows that self-detection increases the total expected QALYs, while reducing the number of mammograms. This POMDP is solved optimally using Monahan's algorithm; however their results are sensitive to the disutility values associated with a mammogram.

Models have also been used to optimize cancer screening and treatment decisions for colorectal cancer. For example, *Erenay et al.* (2014) developed a POMDP to optimize colonoscopy screening policies to maximize expected QALYs. They report

that optimal screening policies recommend women with a history of colorectal cancer should be screened via colonoscopy more frequently than men, while women without a history of colorectal cancer should be screened via colonoscopy less frequently than men.

2.3 Contributions to the Literature

Zhang et al. (2012a) and *Underwood et al.* (2012) developed models to study optimal prostate cancer screening policies that maximize expected QALYs; however, both of these models only have one preclinical cancer state. Approximately half of screen-detected cases are considered low-risk with slow growing tumors, and these patients have different treatment options than patients with high-risk tumors. Thus, our work extends that of *Zhang et al.* (2012a) and *Underwood et al.* (2012) by accounting for different grades of prostate cancer, and can therefore differentiate between different treatment options for patients. Our simulation models in Chapters III and IV also contribute to the literature by evaluating new prostate cancer diagnostic biomarkers and MRI, including multi-stage approaches for using these new technologies. Our work provides a way to directly compare the long-term health impacts of these new technologies. Additionally, we have included behavioral aspects of treatment choice into our simulation models by using a predictive model for active surveillance selection. In Chapter V, we developed a POMDP, which we solved using a new data-driven sampling approach to develop a set of *relevant* grid points in the belief space based on our special problem structure. Finally, in Chapter VI we present what is, to the best of our knowledge, the first hidden Markov model based on longitudinal active surveillance data. We fit our model using data from 1499 patients enrolled in active surveillance at Johns Hopkins over 20 years. Using this hidden Markov model, we quantified the trade-off among a large number of biopsy strategies and provide insights about how to improve upon strategies proposed in the literature.

CHAPTER III

A Simulation Model of Prostate Cancer Screening Decisions Using Newly Discovered Biomarkers

3.1 Introduction

Although prostate cancer is the most common solid tumor in American men, there is controversy surrounding prostate cancer screening. The AUA recommends shared decision-making for men from ages 55 to 69 considering PSA-based screening, and specifies screening intervals of two years preserve the majority of the benefits of screening and reduce overdiagnosis and false positives (*Carter et al.* (2013)). However, the U.S. Preventive Services Task Force recommends against prostate cancer screening with the PSA test due to the potential harms from unnecessary biopsies and overtreatment of low-risk disease (*Moyer* (2012)). In recent years many new biomarkers have been discovered for early detection of prostate cancer that may be able to supplement the PSA test to reduce unnecessary biopsies. Patients and their healthcare providers now have access to these new biomarkers which could potentially be combined into multi-stage biomarker screening strategies that improve the precision with which screening can be performed. These discoveries have the potential to improve patient survival and lower the burden of screening by better discriminating between patients with and without cancer. However, these tests vary in their predic-

tive characteristics, and the ideal way to use them to achieve optimal long-term health benefits is unclear. In this chapter we study the question of how to design *two-stage biomarker screening strategies* in the context of prostate cancer. A two-stage strategy uses one biomarker (e.g. PSA) to stratify patients into two groups that either receive biopsy or no biopsy and a third group that receives a second stage biomarker test. Such strategies have to potential to better select men for biopsy.

Several new diagnostic prostate cancer biomarkers have recently come to market (*Makarov et al.* (2009); *Tosoian et al.* (2016)). Some of these biomarkers are PSA derivatives, such as free PSA and [-2]proPSA. Some of the biomarkers are based on combinations of serum markers, such as phi, which uses a combination of total PSA, free PSA, and [-2]proPSA to generate a score (*Bryant and Lilja* (2014); *Catalona et al.* (2011)), and the 4Kscore, which uses a panel of total PSA, free PSA, intact PSA, and hK2 to estimate a patient's risk of high-grade cancer (Gleason score ≥ 7) on biopsy. Other molecular biomarkers include PCA3 and T2:ERG, which are detectable in post-DRE urine (*Bussemakers et al.* (1999); *Salagierski and Schalken* (2012); *Truong et al.* (2013); *Tomlins et al.* (2005); *Brenner et al.* (2013); *Pettersson et al.* (2012); *Young et al.* (2012)). The MiPS early detection test combines a patient's serum PSA, urine PCA3 score, and urine T2:ERG score into a single multivariate regression model to estimate individualized risk estimates for all prostate cancer and high-grade prostate cancer (*Tomlins et al.* (2016)). These tests vary in the outcome they predict (all-cancer versus high-grade cancer) and in their sensitivities and specificities. No study has yet attempted to compare these biomarkers to determine which characteristics achieve optimal long-term health outcomes in the context of early detection of prostate cancer.

In order to better understand the optimal design of screening strategies in a multi-biomarker setting, we estimated long-term health outcomes using a partially observable Markov model. We validated the model by comparing model-based estimates of

health outcomes with independent estimates reported in the literature. Based on the AUA screening policy, we compared each of the biomarkers in the context of patients who were screened from ages 55 to 69 with a screening interval of two years. During each screening period, we employed an innovative two-stage biomarker screening strategy. If the patient's serum PSA was over a specified threshold (2 or 4 ng/mL), a second biomarker test was administered. We estimated the number of prostate cancer deaths and screening biopsies per 1000 men, as well as the gain in QALYs compared to no screening in order to identify the ideal biomarker characteristics. We drew conclusions about optimal screening strategy design characteristics that may generalize to other disease contexts in which multiple biomarkers can be used to achieve early detection.

3.2 Model

To evaluate screening strategies that use biomarkers of varying sensitivity and specificity, we developed a partially observable Markov model in which pretreatment states are not directly observable. Biomarker tests give (imperfect) information about the true state of the patient. The partially observable pretreatment states in the model include no prostate cancer, undetected organ-confined prostate cancers based on Gleason score (GS $<$ 7, GS $=$ 7, GS $>$ 7), and extraprostatic or lymph node-positive cancer (EPLN). The EPLN state aggregates these two conditions into one state because they are similarly associated with decreased survival. The states were selected because they distinguish patients on the basis of likely treatment options, outcomes, and survival.

3.2.1 Model Parameters

Figure 3.1 displays the health states and possible state transitions for the model. Each year that the screening strategy calls for testing the following sequence of events

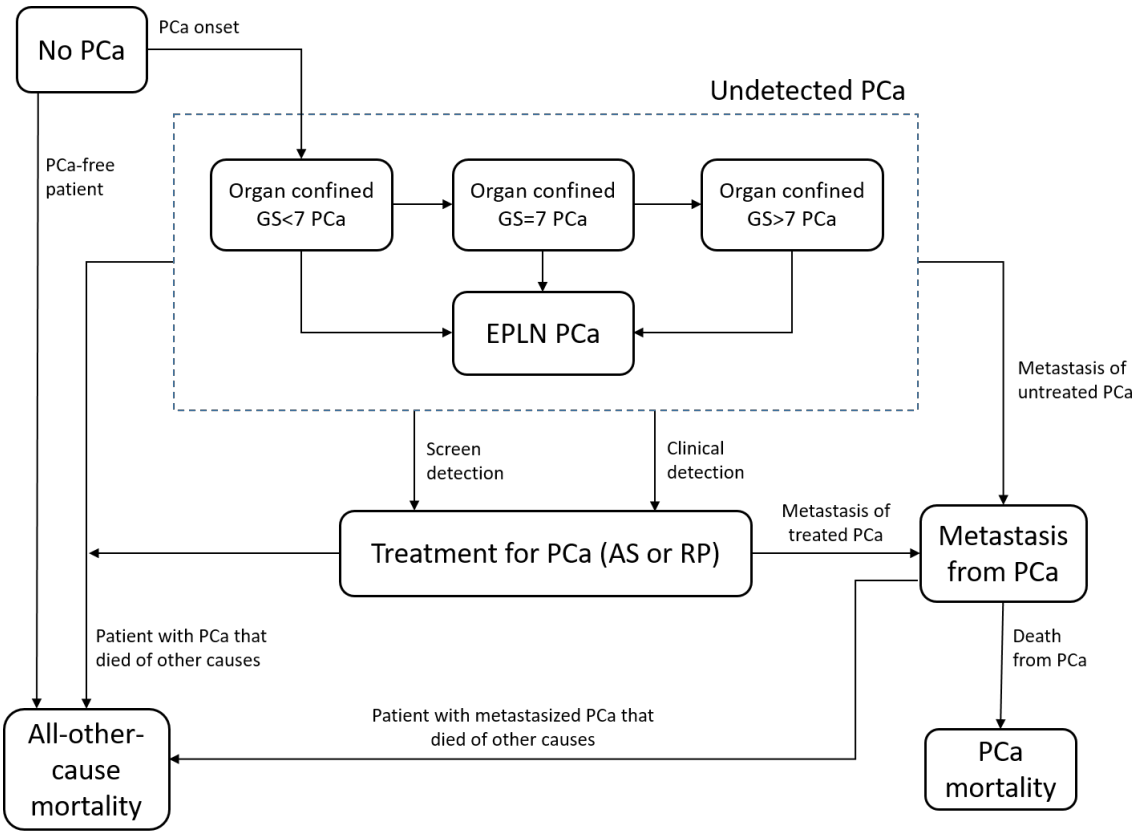

Figure 3.1: State transition diagram. Health states and progression paths in the Markov model are shown, where transitions between states are represented by arrows. Patients who are detected with prostate cancer (PCa) are treated immediately with radical prostatectomy (RP) or active surveillance (AS). GS = Gleason score; EPLN = extraprostatic or lymph node-positive cancer.

in the model occur: the patient receives one or more biomarker tests according to the specified strategy; the biomarker test results determine whether a biopsy is performed; and the patient transitions to their next health state. As our model focuses on screening of the general population, the screening strategy terminates after an initial biopsy and the patient continues to make state transitions in the absence of screening until reaching one of the absorbing states, all-other-cause mortality or prostate cancer mortality. The parameters used to calculate the transition probabilities are described in Table 3.1, and how these parameters were calculated is described in this section.

Table 3.1: Parameters, their sources, and the specific values used in our base case and sensitivity analysis.

Parameter	**Symbol**	**Low Value(s)**	**Base Case Value(s)**	**High Value(s)**	**Source**
Annual transition rate from No PCa to GS < 7	w_t	Lower bound of 95% C.I.	0.004–0.069	Upper bound of 95% C.I.	*Haas et al.* (2007)
Annual other-cause mortality rate	d_t	-20%	0.002–0.347	+20%	*Arias* (2010)
Annual metastasis rate for patients with undiagnosed PCa	e_t	-10%	0.002–0.035	+10%	Calibrated
Annual PCa-specific mortality rate given metastasized PCa	z_t	-10%	0.181–0.204	+10%	*Ries et al.* (2007)
Sensitivity of prostate biopsy procedure	f	-10%	0.8	+10%	*Haas et al.* (2007)
Annual transition rate from GS < 7 to GS $= 7$	$o1o2$	-10%	0.101	+10%	*Draisma et al.* (2003)
Annual transition rate from GS $= 7$ to GS > 7	$o2o3$	-10%	0.087	+10%	*Draisma et al.* (2003)
Annual transition rate from GS < 7 to EPLN	$o1e$	-10%	0.029	+10%	*Draisma et al.* (2003)
Annual transition rate from GS $= 7$ to EPLN	$o2e$	-10%	0.081	+10%	*Draisma et al.* (2003)
Annual transition rate from GS > 7 to EPLN	$o3e$	-10%	0.097	+10%	*Draisma et al.* (2003)
Probability of no possible recurrence following definitive treatment in state EPLN	pnc	-10%	0.468	+10%	*Roehl et al.* (2004)
Proportion of patients detected with GS < 7 who undergo active surveillance	s	-10%	0.485	+10%	*Liu et al.* (2015)
Annual metastasis rate for patients with possible recurrence after definitive treatment in EPLN	g	-10%	0.006	+10%	Mayo Clinic Radical Prostatectomy Registry
Instantaneous QALY disutility for screening	δ_{Scr}	0.0	0.00019	0.00019	*Heijnsdijk et al.* (2012)
Instantaneous QALY disutility for a prostate biopsy	δ_{Biop}	0.00346	0.00577	0.00750	*Heijnsdijk et al.* (2012)
Instantaneous QALY disutility for PCa diagnosis	δ_{Dia}	0.0125	0.01667	0.0208	*Heijnsdijk et al.* (2012)
Instantaneous QALY disutility for radical prostatectomy	δ_{Tre}	0.0917	0.24667	0.323	*Heijnsdijk et al.* (2012)
Annual QALY disutility for 9-year post-radical prostatectomy recovery period	δ_{Rec}	0.0	0.05	0.07	*Heijnsdijk et al.* (2012)
Annual QALY disutility for active surveillance	δ_{AS}	0.0	0.03	0.15	*Heijnsdijk et al.* (2012)
Annual QALY disutility for metastasis	δ_{Met}	0.14	0.4	0.76	*Heijnsdijk et al.* (2012)

PCa = prostate cancer; GS = Gleason score; C.I. = confidence interval; EPLN = extraprostatic or lymph node-positive cancer; QALY = quality-adjusted life year.

Annual transition rate from No PCa to GS < 7 (w_t); We assume that when patients are in No PCa, they can only transition to organ confined GS < 7 or death, which is consistent with the assumption made in *Draisma et al.* (2003). In Figure 2 of *Haas et al.* (2007), the predicted proportion of patients with prostate cancer is reported by age based on needle biopsies on autopsy prostates. Let w_t be the annual transition probability from No PCa to GS < 7 during the period from t to $t+1$. Let a_t and a_{t+1} be the predicted proportion of patients with prostate cancer at age t and age t+1, respectively. Let N_t and N_{t+1} denote the size of the male population at age t and $t+1$, respectively. Then, w_t is the cumulative incidence over the age range t to $t+1$, and we have the following relation:

$$a_{t+1}N_{t+1} = a_t N_t \frac{N_{t+1}}{N_t} + (1-a_t) w_t N_t \frac{N_{t+1}}{N_t}$$

where $a_{t+1}N_{t+1}$ is the number of prostate cancer patients at age $t+1$, $a_t N_t \frac{N_{t+1}}{N_t}$ is the number of prostate cancer patients who developed prostate cancer at age t and are still alive at age $t+1$, and $(1-a_t)w_t N_t \frac{N_{t+1}}{N_t}$ is the number prostate cancer patients at age $t+1$ who did not have prostate cancer at age t. Note that in a closed population, $N_t > N_{t+1}$. Thus, the ratio $\frac{N_{t+1}}{N_t}$ is the proportion of age t patients who live to age $t+1$. By simple algebraic manipulation, we can simplify the equation to:

$$w_t = \frac{a_{t+1} - a_t}{1 - a_t}$$

We used this equation to calculate the annual transition rates based on the estimates of a_t reported in *Haas et al.* (2007).

Annual other-cause mortality rate (d_t); The annual other-cause mortality rate was obtained from the CDC Life Tables (*Arias* (2010)).

Annual metastasis rate for patients with undiagnosed prostate cancer (e_t); Since the metastasis rate for patients with undiagnosed prostate cancer is un-

observed, we calibrated the values of e_t with respect to long-term outcomes reported in the literature. Specifically, we varied the metastasis rate in 10-year periods, and calibrated the values so the resulting age-dependent risk of prostate cancer death under routine screening matched the values reported in the literature (*Howlader et al.* (2012)):

Age	**Risk of prostate cancer death**
50	2.82%
60	2.98%
70	3.18%
80	3.36%

For calibration, 30,000,000 samples were taken and confidence intervals were 0.01%. The resulting age-dependent metastasis rates were:

t	e_t
40–59	0.0017
60–69	0.0054
70–79	0.0139
80–89	0.0343
90–100	0.0345

Annual prostate cancer-specific mortality rate given metastasized prostate cancer (z_t); Let S_t be defined as the five-year prostate cancer-specific survival rate for metastasized prostate cancer at age t. Then we can calculate annual prostate cancer-specific mortality rate given metastasized prostate cancer, z_t, as:

$$z_t = 1 - S_t^{1/5}$$

Table 22.5 of *Ries et al.* (2007) reports the 5-year prostate cancer-specific mortality rate for metastasized prostate cancer to be 0.319, 0.366, and 0.368 for age groups 20–64, 65–74, and 75+, respectively. Thus,

$$z_t = \begin{cases} 0.204, & \text{if } t \leq 64 \\ 0.182, & \text{if } 65 \leq t \leq 74 \\ 0.181, & \text{if } t \geq 75 \end{cases}$$

Sensitivity of prostate biopsy procedure (f); *Haas et al.* (2007) reported the sensitivity of biopsy from the mid peripheral zone and the lateral peripheral zone combined for clinically significant cancer to be 80%.

Annual transition rate calculations from *Draisma et al.* (2003); The model in *Draisma et al.* (2003) has three localized stages (Loc G1, Loc G2, Loc G3). Loc G1 is equivalent to organ confined GS < 7 PCa in our model, Loc G2 is equivalent to organ confined GS $= 7$ PCa in our model, and Loc G3 is equivalent to organ confined GS > 7 PCa in our model. The model in *Draisma et al.* (2003) has three regional stages (Reg G1, Reg G2, Reg G3), and our model has one regional state (EPLN). *Draisma et al.* (2003) reported the average dwelling time (μ) for each state. We used these reported dwelling times to calculate stationary transition probabilities between states. For a discrete-time Markov chain, we let T_i be the time spent in state i before transitioning to another state. T_i is a random variable of geometric distribution, which has a mean of $\frac{1}{p}$ where p is the annual transition rate of leaving the state. Thus, we can use μ to calculate the annual transition rate of leaving a state: $\frac{1}{\mu}$.

- **Annual transition rate from GS < 7 to GS $= 7$ ($o1o2$);** *Draisma et al.* (2003) reported the average dwelling time in Loc G1 to be 6.95, and reported that 0.7 proportion of patients departing Loc G1 transition to Loc G2. Thus the transition rate of leaving the state of GS < 7 is $\frac{1}{6.95} = 0.144$, and the transition rate from GS < 7 to GS $= 7$ is $0.144 \times 0.70 = 0.101$.

- **Annual transition rate from GS = 7 to GS > 7 ($o2o3$);** *Draisma et al.* (2003) reported the average dwelling time in Loc G2 to be 4.81, and reported that 0.42 proportion of patients departing Loc G2 transition to Loc G3. Thus the transition rate of leaving the state of GS = 7 is $\frac{1}{4.81} = 0.208$, and the transition rate from GS = 7 to GS > 7 is $0.208 \times 0.42 = 0.087$.

- **Annual transition rate from GS < 7 to EPLN ($o1e$);** *Draisma et al.* (2003) reported the average dwelling time in Loc G1 to be 6.95, and reported that 0.2 proportion of patients departing Loc G1 transition to regional disease. Thus the transition rate of leaving the state of GS < 7 is $\frac{1}{6.95} = 0.144$, and the transition rate from GS < 7 to EPLN is $0.144 \times 0.20 = 0.029$.

- **Annual transition rate from GS=7 to EPLN ($o2e$);** *Draisma et al.* (2003) reported the average dwelling time in Loc G2 to be 4.81, and reported that 0.39 proportion of patients departing Loc G2 transition to regional disease. Thus the transition rate of leaving the state of GS = 7 is $\frac{1}{4.81} = 0.208$, and the transition rate from GS = 7 to EPLN is $0.208 \times 0.39 = 0.081$.

- **Annual transition rate from GS > 7 to EPLN ($o3e$);** *Draisma et al.* (2003) reported the average dwelling time in Loc G3 to be 5.25, and reported that 0.51 proportion of patients departing Loc G3 transition to regional disease. Thus the transition rate of leaving the state of GS > 7 is $\frac{1}{5.25} = 0.190$, and the transition rate from GS > 7 to EPLN is $0.190 \times 0.51 = 0.097$.

Metastasis of treated prostate cancer (pnc, g); Metastasis following radical prostatectomy depends on the stage of the disease at treatment. There are two post-treatment states patients can transition to following treatment: no recurrence following treatment (NRFT) and possible recurrence following treatment (PRFT). If a patient has organ-confined disease at surgery, they transition directly to NRFT following radical prostatectomy. Patients who transition to NRFT have been cured

and will not develop metastasis. If a patient has extraprostatic or lymph node-positive disease at treatment, they transition to NRFT with probability 0.468 (defined as *pnc* in Table 3.1), and they transition to PRFT with probability 0.532 (*Roehl et al.* (2004)). The annual metastasis rate for patients in PRFT is 0.006 based on the Mayo Clinic Radical Prostatectomy Registry (defined as g in Table 3.1) as used in a previous prostate cancer model (*Zhang et al.* (2012a)). If a patient's Gleason score was upgraded as a result of a surveillance biopsy, they were assumed to have a radical prostatectomy. From the post-diagnosis states patients eventually transition to metastasis and/or death from prostate cancer or other causes.

Proportion of patients detected with GS < 7 who undergo active surveillance (s); We estimated the mean probability that low-risk patients initiate active surveillance to be 0.485 based on the logistic regression model presented in *Liu et al.* (2015).

QALY Disutilities; The QALY disutility values were all obtained from *Heijnsdijk et al.* (2012). As an example calculation, *Heijnsdijk et al.* (2012) reported a 0.9 utility for 3 weeks following a prostate biopsy. This is equivalent to a 0.9 utility for 3 weeks, and a 1.0 utility for the remaining 49 weeks of the year. Thus, the annual utility in a year where the patient receives a prostate biopsy is:

$$0.9 \times \frac{3}{52} + 1.0 \times \frac{49}{52} = 0.99423$$

and the QALY disutility for the year is $1 - 0.99423 = 0.00577$. We calculated the other disutilities using the same technique. The disutility for living with metastasis is based on the disutility of palliative therapy presented in *Heijnsdijk et al.* (2012).

3.2.2 Prostate Cancer Screening

The structure of the two-stage biomarker screening strategy is illustrated in Figure 3.2 in which two thresholds divide PSA values into low, intermediate, and high. A patient receives a biopsy if his PSA value is high ($>$ 10 ng/mL). If his PSA value is low at a given screening age, then no biopsy is recommended. If the PSA is between the low and high thresholds, then a second biomarker test is employed. If the second biomarker test is positive, the patient receives a biopsy; otherwise, the patient does not receive a biopsy and continues to be screened in future years. We evaluated two PSA thresholds to trigger a second biomarker test: 2 and 4 ng/mL. We selected these thresholds because it has been reported that phi, 4Kscore, and [-2] proPSA have the ability to select men with PSA values of 2-10 ng/mL for prostate biopsy, and because 4 ng/mL is a commonly used biopsy threshold (*Bratt and Lilja* (2015)). We chose to use this two-stage screening strategy for multiple reasons. First, PSA is an established test and many new biomarkers are only approved to be used along with the PSA test. Second, new biomarkers can be expensive, and this approach pragmatically uses the new biomarkers when they will add greatest value and does not use them when they have little value. Additionally, we assumed 100% adherence to the screening strategy for our base case, and performed sensitivity analysis on the adherence rates.

We sampled PSA scores using a random effects model that includes the patient's current age and their age at onset of a preclinical tumor (*Gulati et al.* (2010)):

$$\log\{y_i(t)\} = \beta_{0i} + \beta_{1i}t + \beta_{2i}(t - t_{oi})I(t > t_{oi}) + \epsilon,$$

where $y_i(t)$ is the PSA level for individual i at age t, $t = 0$ corresponds to age 35, t_{oi} is the age at onset of a preclinical tumor for individual i, I is an indicator function, and ϵ is random noise. Finally, individual intercepts and slopes for each individual i are given by $\beta_{ki} \sim N(\mu_k, \sigma_k^2)$ for $k = 0, 1, 2$.

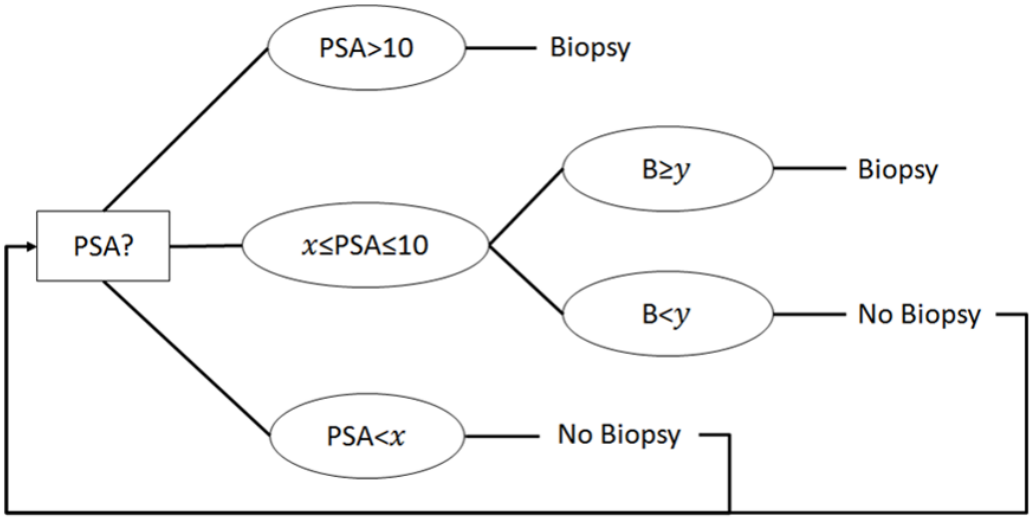

Figure 3.2: Two-stage biomarker screening strategy where the result of the PSA test determines whether a second biomarker is used. If a patient's PSA score is greater than 10 ng/mL, they will automatically receive a biopsy. B represents the observed second biomarker result for the patient, x is the PSA threshold to trigger a second biomarker test, and y is the threshold for the second biomarker to trigger biopsy.

Table 3.2: Biomarker sensitivities and specificities for all-cancer and high-grade cancer (Gleason score ≥ 7) reported in the literature. The sensitivities and specificities for 4Kscore and the MiPS tests were calculated using data presented in *Parekh et al.* (2015) and from the study *Tomlins et al.* (2016), respectively. These 14 tests were evaluated for two PSA thresholds (2 ng/mL and 4 ng/mL), resulting in 28 screening strategies. Blank entries for thresholds indicate no threshold given in the source.

Biomarker Test	**Threshold**	**Sensitivity**	**Specificity**	**Source**
All-cancer				
% p2PS	≥ 1.7	0.70	0.70	*Ferro et al.* (2012)
% p2PS	≥ 2.5	0.38	0.90	*Ferro et al.* (2012)
phi	≥ 38.7	0.85	0.61	*Ferro et al.* (2012)
PCA3	–	0.93	0.37	*Salami et al.* (2013)
T2:ERG	–	0.67	0.87	*Salami et al.* (2013)
T2:ERG	–	0.37	0.93	*Sartori and Chan* (2014)
High-grade cancer				
4Kscore	$\geq 9\%$	0.90	0.52	*Parekh et al.* (2015)
4Kscore	$\geq 12\%$	0.86	0.62	*Parekh et al.* (2015)
4Kscore	$\geq 15\%$	0.79	0.70	*Parekh et al.* (2015)
All-cancer MiPS	$\geq 25\%$	0.94	0.41	*Tomlins et al.* (2016)
All-cancer MiPS	$\geq 52\%$	0.68	0.78	*Tomlins et al.* (2016)
High-grade MiPS	$\geq 10\%$	0.95	0.36	*Tomlins et al.* (2016)
High-grade MiPS	$\geq 15\%$	0.88	0.55	*Tomlins et al.* (2016)
High-grade MiPS	$\geq 26\%$	0.70	0.76	*Tomlins et al.* (2016)

For the sensitivity and specificity of the second biomarker test, we used values reported in the literature. We performed a systematic review of the literature and chose the sensitivities and specificities that were non-dominated (i.e., biomarkers such that no other biomarker had both a higher sensitivity and a higher specificity). Table 3.2 shows sensitivity and specificity values we used for all-cancer and high-grade cancer (Gleason score ≥ 7). These 14 (second-stage) tests were evaluated for two (first-stage) PSA thresholds (2 ng/mL and 4 ng/mL), resulting in 28 screening strategies. Biopsy results were randomly sampled as either positive or negative, assuming a sensitivity of 0.8 (*Haas et al.* (2007)). If the biopsy result was positive, we estimated the probability that the biopsy provides an incorrect grading at diagnosis based on data reported in *Epstein et al.* (2012).

3.2.3 Clinical Detection of Prostate Cancer

Patients were diagnosed with prostate cancer in one of two ways: by routine screening (i.e., an elevated biomarker score that leads to a positive biopsy) or by clinical detection (i.e., prostate cancer that develops symptoms). We assumed that the lead time clock for clinical detection starts once a patient has both prostate cancer and a PSA score ≥ 3 ng/mL. *Savage et al.* (2010) developed a distribution of lead times from an elevated PSA measurement of ≥ 3 ng/mL to clinical diagnosis of prostate cancer. For each patient, we randomly sampled a lead time from this distribution. If a patient's lead time is x years, after the patient has had prostate cancer and a PSA score ≥ 3 ng/mL for x years, if the patient is alive and has neither been diagnosed nor treated for prostate cancer, then the patient is assumed to be clinically detected.

3.2.4 Prostate Cancer Treatment

Following diagnosis, patients received watchful waiting, active surveillance or radical prostatectomy. We assumed patients with Gleason score ≥ 7 received radical prostatectomy. Patients diagnosed with Gleason score < 7 were assumed to be treated via active surveillance or radical prostatectomy. Based on practice patterns reported in *Liu et al.* (2015), we assumed that 48.5% of patients diagnosed with Gleason score < 7 received active surveillance, while the other 51.5% received radical prostatectomy. Given the lack of consensus in published guidelines for active surveillance, we assumed that patients received a biopsy one year after diagnosis, followed by a biopsy every two years for 10 years following diagnosis (*Cooperberg et al.* (2011)). Patients over age 80 were assumed to receive watchful waiting.

Patients receiving active surveillance continue to progress through the natural history of the disease until they have a biopsy result of Gleason score ≥ 7. We made the same assumptions about surveillance biopsies as described above. If a patient's

Gleason score was upgraded as a result of a surveillance biopsy, they were assumed to have a radical prostatectomy. However, if they are never detected to have higher risk disease, they have the survival of an untreated patient. Survival following radical prostatectomy depends on the stage of the disease at treatment. There are two post-treatment states patients can transition to following treatment: no recurrence following treatment (NRFT), and possible recurrence following treatment (PRFT). If a patient has organ-confined disease at surgery, they transition directly to NRFT. If a patient has extraprostatic or lymph node-positive disease at treatment, they transition to NRFT with probability 0.468 (defined as *pnc* in Table 3.1), and they transition to PRFT with probability 0.532. The annual metastasis rate for patients in PRFT is 0.006 based on the Mayo Clinic Radical Prostatectomy Registry (defined as g in Table 3.1). From the post-diagnosis states patients eventually transition to metastasis and/or death from prostate cancer or other causes.

3.2.5 Model Validation

To perform model validation, we compared estimates of clinical statistics from our model with literature estimates. The model estimates were based on the assumption that all men were screened annually from age 50 to 75 with a PSA threshold of 4 ng/mL, because that was a common strategy at the time upon which the literature estimates are based (*Ross et al.* (2000); *Andriole et al.* (2009)). We compared our model results with independent estimates from the literature for age-dependent risks of prostate cancer death, expected lifespan for a 40-year-old man, age-dependent risks of prostate cancer diagnosis, Gleason score distribution at diagnosis, and biopsy-detectable prostate cancer prevalence rates by age.

3.2.6 Simulation Parameters

The AUA recommends shared decision-making for men considering PSA-based screening from ages 55 to 69 with a screening interval of two years. Based on this recommendation, patients were PSA-screened every two years from ages 55 to 69 (*Carter et al.* (2013)). Each patient simulation began at age 40. The model was used to evaluate 28 different prostate cancer screening strategies based on published estimates of sensitivity and specificity for biomarkers reported in the literature. Table 3.2 shows the sensitivity and specificity values for all cancer and high-grade cancer (Gleason score ≥ 7). We compared these values with using PSA alone and to hypothetical perfect biomarkers that have a sensitivity and specificity of 1.0 for either all cancer or high-grade cancer. We also investigated the trade-off of sensitivity and specificity by evaluating long-term health outcomes for patients under 30 different thresholds for the high-grade MiPS test. To perform this analysis, we used a large data set of PSA, PCA3, and T2:ERG scores from a presumed cancer-free population of patients undergoing diagnostic prostate biopsy to estimate the high-grade sensitivity and specificity of the high-grade MiPS test under each threshold (*Tomlins et al.* (2016)).

For each strategy evaluated, we estimated the mean number of screening biopsies and prostate cancer deaths per 1000 men, and the mean QALYs gained per 1000 men relative to no screening. Our QALY measurements account for disutilities of screening, biopsy, diagnosis, active surveillance, radical prostatectomy, recovery from radical prostatectomy, and metastasis; the values of the disutilities with their sources are shown in Table 3.1. The reward update function for QALYs was:

$$r_t(s_t, a_t) = 1 - \delta_{\text{Scr}}(a_t) - \delta_{\text{Biop}}(a_t) - \delta_{\text{Dia}}(a_t) - \delta_{\text{Tre}}(a_t) - \delta_{\text{Rec}}(a_t) - \delta_{\text{AS}}(a_t) - \delta_{\text{Met}}(s_t)$$

where $r_t(s_t, a_t)$ is the reward a patient receives at age t, which is 1 minus the disutilities associated with screening, biopsy, diagnosis, treatment and the presence of metastatic

cancer, as defined in Table 3.1. The arguments for the reward are the health state, s_t, that defines the cancer status of the patient and the action, a_t, that defines whether a screening test or biopsy was performed. The total expected QALYs a patient receives in their lifetime is:

$$R = \mathbb{E}^{\pi}\left[\sum_{t=40}^{T} r_t(s_t, a_t)\right] \tag{3.1}$$

where T denotes maximum lifespan and the expectation is with respect to the stochastic process induced by the screening strategy π that defines the frequency of testing and the thresholds at which to perform biomarker tests and/or biopsies. This amounts to assuming a risk neutral decision maker (e.g. the patient). Since we are not analyzing costs, we did not use a discount factor. Since exact evaluation of R in equation 3.1 is not straight-forward due to history dependence of rewards up to a given decision epoch t, we used forward simulations to obtain statistical estimates based on N patient samples each starting at age 40:

$$\hat{R} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=40}^{T} r_t^{(n)}(s_t, a_t) \tag{3.2}$$

We synchronize patient histories using the method of common random numbers. Each patient has his own stream of random numbers to determine his sequence of health states and test results. This approach allows our model to compare a patient's "natural history" of prostate cancer in the absence of screening to their health outcomes under several screening approaches.

Simulation was performed to generate sample paths and obtain statistical estimates of expected rewards for each strategy. This simulation model was implemented in C/C++. We ran each strategy for 30,000,000 sample paths, which took less than 12.5 minutes to run using 3.40 GHz with 16 GB of RAM. The largest 95% confidence interval reflecting Monte Carlo error was less than 1% of the corresponding sample-mean point estimate.

3.2.7 Sensitivity Analysis

We performed one-way sensitivity analysis on all of the model parameters by estimating $\hat{R}$ for varying choices of each parameter from a low to a high value, as defined in Table 3.1. We also performed probabilistic sensitivity analysis, during which we varied each model parameter by sampling from a uniform distribution between the low and high values reported in Table 3.1. During the probabilistic sensitivity analysis, we performed 30 experiments with 30,000,000 sample paths for each experiment. Additionally, we looked at the impact of varying screening adherence (i.e., participation and attendance rates). We looked at the effect of varying these parameters on the expected number of prostate cancer deaths per 1000 men and the increase in QALYs per 1000 men relative to no screening. To perform sensitivity analysis, we used the strategy with a PSA threshold of 2 ng/mL and a second biomarker test with a high-grade sensitivity and specificity of 0.86 and 0.62, respectively.

3.3 Results

3.3.1 Model Validation

Table 3.3 compares estimates of clinical statistics from our model with literature estimates from external validation studies. Overall, our estimates from the model compare well with estimates from the literature. The SEER estimates that we have compared to in Table 3.3 are from the years 2006 to 2008 (*Howlader et al.* (2012)). Any variations are most likely due to our assumption that patients have perfect adherence to the screening strategy.

Table 3.4 presents validation results in the absence of screening. Compared to Table 3.3, the risk of prostate cancer death is higher, expected lifespan is lower, and the Gleason score distribution at diagnosis shifts to higher grade disease. The risk of diagnosis decreases for younger ages. After age 80, patients with advanced stage

Table 3.3: Results from validation of the Monte Carlo simulation model based on the partially observable Markov chain. The model estimates were based on the assumption that all men were screened for prostate cancer (PCa) annually from age 50 to 75 with a PSA threshold of 4 ng/mL.

Statistic	**Model Estimate**		**Literature Estimate**		**Literature Source**
Overall risk of PCa death for 40-year-old man	2.73%		2.73%		*Howlader et al.* (2012)
Age-dependent risk of PCa death	**Age**	**Risk**	**Age**	**Risk**	
	50	2.82%	50	2.82%	
	60	2.98%	60	2.98%	*Howlader et al.* (2012)
	70	3.18%	70	3.18%	
	80	3.36%	80	3.36%	
Expected lifespan for 40-year-old man (yr.)	38.19		37.7		*Arias* (2010)
Overall diagnosis risk for 40-year-old man	16.5%		16.6%		*Howlader et al.* (2012)
Age-dependent risk of being diagnosed with PCa within 10 years	**Age**	**Risk**	**Age**	**Risk**	
	50	3.1%	50	2.3%	
	60	6.1%	60	6.6%	*Howlader et al.* (2012)
	70	7.0%	70	8.2%	
	80	7.3%	80	5.1%	
Gleason score (GS) distribution at diagnosis	**GS**	**Proportion**	**GS**	**Proportion**	
	< 7	53%	< 7	49%	
	$= 7$	31%	$= 7$	29%	*Draisma et al.* (2003)
	> 7	16%	> 7	22%	
Biopsy-detectable PCa prevalence	**Age**	**Prevalence**	**Age**	**Prevalence**	
	50	13%	50	13%	
	60	22%	60	22%	*Haas et al.* (2007)
	70	36%	70	36%	
	80	50%	80	51%	

disease present with symptoms or develop metastasis, which leads to a large number of delayed diagnoses.

3.3.2 Base Case Analysis

We estimated the expected number of QALYs gained per 1000 men relative to no screening for each of the biomarkers defined in Table 3.2 as well as two hypothetical perfect biomarkers. Ten of the new biomarkers maximized expected QALY gains with overlapping confidence intervals. The performance outcomes for these ten biomarkers are shown in Table 3.5 along with the results for the hypothetical perfect biomarkers. While there was no statistically significant difference between these ten tests in the QALYs gained per 1000 men, the number of biopsies per 1000 men varied from 184 to 237. These ten tests also performed significantly better than using PSA alone with a

Table 3.4: Results of the Monte Carlo simulation model based on the partially observable Markov chain for the case of no screening for prostate cancer (PCa).

Statistic	**Model Estimate**	
Overall risk of PCa death for 40-year-old man	3.23%	
Age-dependent risk of PCa death	**Age**	**Risk**
	50	3.33%
	60	3.53%
	70	3.80%
	80	4.00%
Expected lifespan for 40-year-old man (yr.)	38.15	
Overall diagnosis risk for 40-year-old man	12.0%	
Age-dependent risk of being diagnosed with PCa within 10 years	**Age**	**Risk**
	50	0.7%
	60	3.1%
	70	7.0%
	80	10.2%
Gleason score (GS) distribution at diagnosis	**GS**	**Proportion**
	< 7	39%
	$= 7$	34%
	> 7	27%
Biopsy-detectable PCa prevalence	**Age**	**Prevalence**
	50	13%
	60	22%
	70	36%
	80	50%

Table 3.5: Best performing strategies in terms of QALYs gained per 1000 men compared to no screening. Each strategy has a PSA threshold of 2 ng/mL to trigger a second biomarker test, and assumes a biopsy will automatically be performed on any patient with a PSA $\geq$ 10 ng/mL.

	Second Biomarker			Expected QALYs	Number of screening	Number of PCa deaths
Test	**Threshold**	**Sensitivity**	**Specificity**	**gained per 1000 men**	**biopsies per 1000 men**	**per 1000 men**
Perfect: HG *	–	1.00	1.00	21.04	128.2	27.5
4Kscore *	$\geq$ 12%	0.86	0.62	18.59	211.9	27.7
4Kscore *	$\geq$ 15%	0.79	0.70	18.52	200.2	27.8
4Kscore *	$\geq$ 9%	0.90	0.52	18.51	222.6	27.6
HG MiPS *	$\geq$ 15%	0.88	0.55	18.48	219.6	27.7
MiPS *	$\geq$ 25%	0.94	0.41	18.38	231.7	27.6
HG MiPS *	$\geq$ 10%	0.95	0.36	18.30	235.0	27.6
Perfect: all	–	1.00	1.00	18.01	146.5	27.1
HG MiPS *	$\geq$ 26%	0.70	0.76	17.93	188.4	27.9
MiPS *	$\geq$ 52%	0.68	0.78	17.79	184.2	28.0
PSA alone	–	–	–	17.75	251.7	27.5
PCA3	–	0.93	0.37	17.65	236.9	27.6
phi	$\geq$ 38.7	0.85	0.61	17.46	218.7	27.6

PCa = prostate cancer; QALYs = quality-adjusted life years.
* Sensitivity and specificity to high-grade (HG) prostate cancer (GS $\geq$ 7).

threshold of 4 ng/mL, achieving between 55% and 65% more QALYs gained per 1000 men. In terms of the initial PSA threshold to trigger a second biomarker test, a PSA threshold of 2 ng/mL performed significantly better than 4 ng/mL in all two-stage strategies, where using an initial PSA threshold of 2 ng/mL achieved between 55% and 65% more QALYs gained per 1000 men than using an initial PSA threshold of 4 ng/mL.

Table 3.6 presents the expected QALYs gained per 1000 men, the number of screening biopsies, and the number of prostate cancer deaths per 1000 men for each of the screening strategies we evaluated. An initial PSA threshold of 2 ng/mL results in more expected QALYs and fewer prostate cancer deaths compared to an initial PSA threshold of 4 ng/mL.

Table 3.6: Strategy performance of all biomarkers in terms of QALYs gained compared to no screening, number of screening biopsies, and number of prostate cancer (PCa) deaths per 1000 men. The results for each PSA threshold are ordered by QALYs gained. Blank entries for thresholds indicate no threshold given in the source.

PSA threshold (ng/mL)	Second Biomarker				Expected QALYs gained per 1000 men	Number of screening biopsies per 1000 men	Number of PCa deaths per 1000 men
	Test	Threshold	Sensitivity	Specificity			
2	Perfect: HG *	–	1	1	21.04	128.2	27.5
	4Kscore *	$\geq 12\%$	0.86	0.62	18.59	211.9	27.7
	4Kscore *	$\geq 15\%$	0.79	0.70	18.52	200.2	27.8
	4Kscore *	$\geq 9\%$	0.9	0.52	18.51	222.6	27.6
	HG MiPS *	$\geq 15\%$	0.88	0.55	18.48	219.6	27.7
	MiPS *	$\geq 25\%$	0.94	0.41	18.38	231.7	27.6
	HG MiPS *	$\geq 10\%$	0.95	0.36	18.30	235.0	27.6
	Perfect: all	–	1	1	18.01	146.5	27.1
	HG MiPS *	$\geq 26\%$	0.70	0.76	17.93	188.4	27.9
	MiPS *	$\geq 52\%$	0.68	0.78	17.79	184.2	28.0
	PSA alone	–	–	–	17.75	251.7	27.5
	PCA3	–	0.93	0.37	17.65	236.9	27.6
	phi	≥ 38.7	0.85	0.61	17.46	218.7	27.6
	%p2PS	≥ 1.7	0.70	0.70	16.69	204.4	27.8
	T2:ERG	–	0.67	0.87	16.66	174.5	27.7
	%p2PS	≥ 2.5	0.38	0.90	14.34	152.2	28.5
	T2:ERG	–	0.37	0.93	14.28	143.4	28.5
4	Perfect: HG *	–	1	1	12.74	104.4	29.2
	4Kscore *	$\geq 9\%$	0.9	0.52	11.71	137.2	29.3
	MiPS *	$\geq 25\%$	0.94	0.41	11.69	142.3	29.3
	4Kscore *	$\geq 15\%$	0.79	0.70	11.68	126.0	29.4
	4Kscore *	$\geq 12\%$	0.86	0.62	11.64	131.7	29.3
	HG MiPS *	$\geq 15\%$	0.88	0.55	11.62	135.5	29.3
	HG MiPS *	$\geq 10\%$	0.95	0.36	11.62	144.2	29.3
	Perfect: all	–	1	1	11.45	113.4	29.1
	HG MiPS *	$\geq 26\%$	0.70	0.76	11.39	120.6	29.5
	MiPS *	$\geq 52\%$	0.68	0.78	11.31	118.8	29.5
	PSA alone	–	–	–	11.27	154.7	29.2
	PCA3	–	0.93	0.37	11.06	145.6	29.2
	phi	≥ 38.7	0.85	0.61	10.87	135.7	29.3
	%p2PS	≥ 1.7	0.70	0.70	10.55	128.0	29.4
	T2:ERG	–	0.67	0.87	10.53	116.8	29.4
	%p2PS	≥ 2.5	0.38	0.90	9.26	105.7	29.8
	T2:ERG	–	0.37	0.93	9.23	102.9	29.8

PCa = prostate cancer; QALYs = quality-adjusted life years.
* Sensitivity and specificity to high-grade (HG) prostate cancer (GS ≥ 7).

Figure 3.3 provides results for the number of screening biopsies and prostate cancer deaths per 1000 men. The figure displays tests from the literature that were on the efficient frontier (i.e., any strategy that resulted in more biopsies and more prostate cancer deaths than another strategy was removed), in addition to the perfect biomarkers and using PSA alone. Figure 3.3 shows the trade-off that occurs between minimizing prostate cancer deaths and minimizing the number of screening biopsies. Although a PSA threshold of 4 ng/mL resulted in fewer biopsies, it also resulted in more prostate cancer deaths. For example, consider the strategy with a PSA threshold of 4 ng/mL and a second biomarker test with a sensitivity and specificity of 0.67 and 0.87 compared to the same strategy with a PSA threshold of 2 ng/mL. The latter strategy is more aggressive, and thus reduces prostate cancer deaths by 6% compared to the former strategy; however, it increases the number of screening biopsies being performed by 49%. As expected, screening strategies with higher sensitivity resulted in fewer prostate cancer deaths and more biopsies, while strategies with higher specificity resulted in fewer biopsies and more prostate cancer deaths. The only two tests that maximized QALYs and also appeared on the efficient frontier of Figure 3.3 was using PSA alone with a threshold of 2 ng/mL and the phi test with threshold of 38.7. Intuitively, using PSA alone with a threshold of 2 ng/mL minimized prostate cancer deaths. The screening strategy that used a PSA threshold of 2 ng/mL and a second biomarker test with high-grade sensitivity and specificity of 0.86 and 0.62, respectively, maximized QALYs, and resulted in a prostate cancer death rate within 1% of using PSA alone with a threshold of 2 ng/mL, while reducing the number of biopsies by 20%.

In addition to the efficient frontier of tests, Figure 3.3 also shows the results for using PSA alone and for hypothetical biomarkers with perfect sensitivity and specificity to all cancer and to high-grade cancer. There exists a two-stage biomarker strategy that can simultaneously reduce the number of prostate cancer deaths and

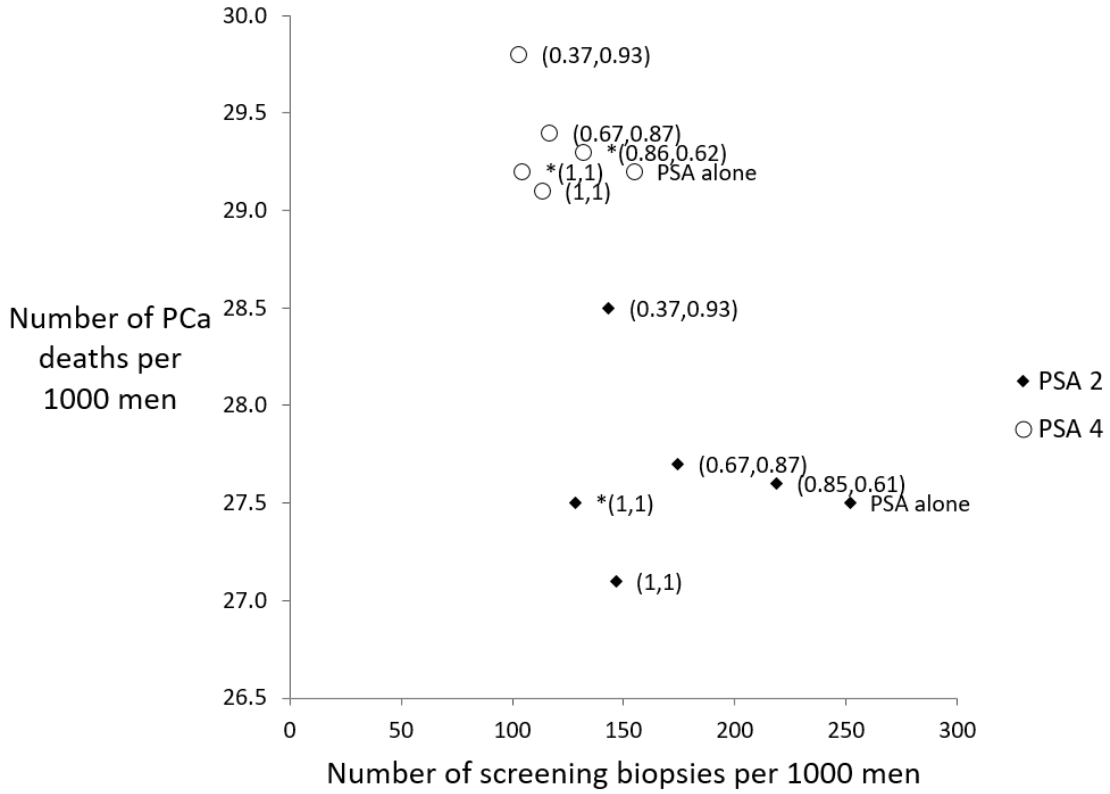

Figure 3.3: Estimated number of prostate cancer (PCa) deaths and screening biopsies per 1000 men from modeled screening strategies. Each point on the graph represents a different screening strategy and is labeled with the sensitivity and specificity of the second biomarker. An asterisk (*) indicates that the sensitivity and specificity are for high-grade prostate cancer (Gleason score ≥ 7). This graph only displays the nondominated strategies of each strategy type, i.e., strategies such that no other strategy resulted in both a lower number of screening biopsies and a lower number of PCa deaths per 1000 men screened (with the exception of the hypothetical perfect biomarkers and PSA alone, which have been shown for reference). The largest 95% confidence interval reflecting Monte Carlo error was less than 1% of the corresponding sample-mean point estimate.

the number of screening biopsies compared to using PSA alone with a threshold of 4 ng/mL. In particular, using a PSA threshold of 2 followed by a test with sensitivity and specificity of 0.37 and 0.93, respectively, can reduce the number of prostate cancer deaths by 2% and the number of screening biopsies by 7% compared to using PSA alone with a threshold of 4 ng/mL. For both PSA thresholds, the test with perfect sensitivity and specificity to high-grade cancer resulted in more prostate cancer deaths but fewer biopsies compared to the test with perfect sensitivity and specificity to all cancer. This further highlights the trade-off between these two competing objectives.

To further investigate the relationship between possible biomarker thresholds, the subsequent sensitivities and specificities that they imply, and long-term health outcomes, we evaluated 30 different thresholds for the high-grade MiPS test using the logistic regression model presented in *Tomlins et al.* (2016); the thresholds we considered ranged from 6% to 35% risk of high-grade cancer on biopsy. Figure 3.4 shows the relationship between the 30 MiPS thresholds, the resulting sensitivity and specificity to high-grade disease, and the mean increase in QALYs per 1000 men compared to no screening. The maximum QALY gain was achieved with a high-grade MiPS threshold of 18% and a corresponding high-grade sensitivity and specificity of 0.83 and 0.63, respectively. Figure 3.4 demonstrates that as specificity is increased and sensitivity is decreased, the expected number of QALYs decreases, which indicates that it is important to maximize sensitivity to high-grade disease in order to maximize expected QALYs. The dotted lines show the 95% confidence interval, showing that thresholds $6 - 29\%$ have confidence intervals that overlap with the maximum achieved at 18%. There is no statistically significant difference in QALYs gained and prostate cancer deaths between the strategies with thresholds from 6% to 22%; however, the number of screening biopsies per 1000 men ranged from 200 to 244. Thus, by looking at performance outcomes in addition to QALYs, we can distinguish between strategies that perform equally well in terms of QALYs.

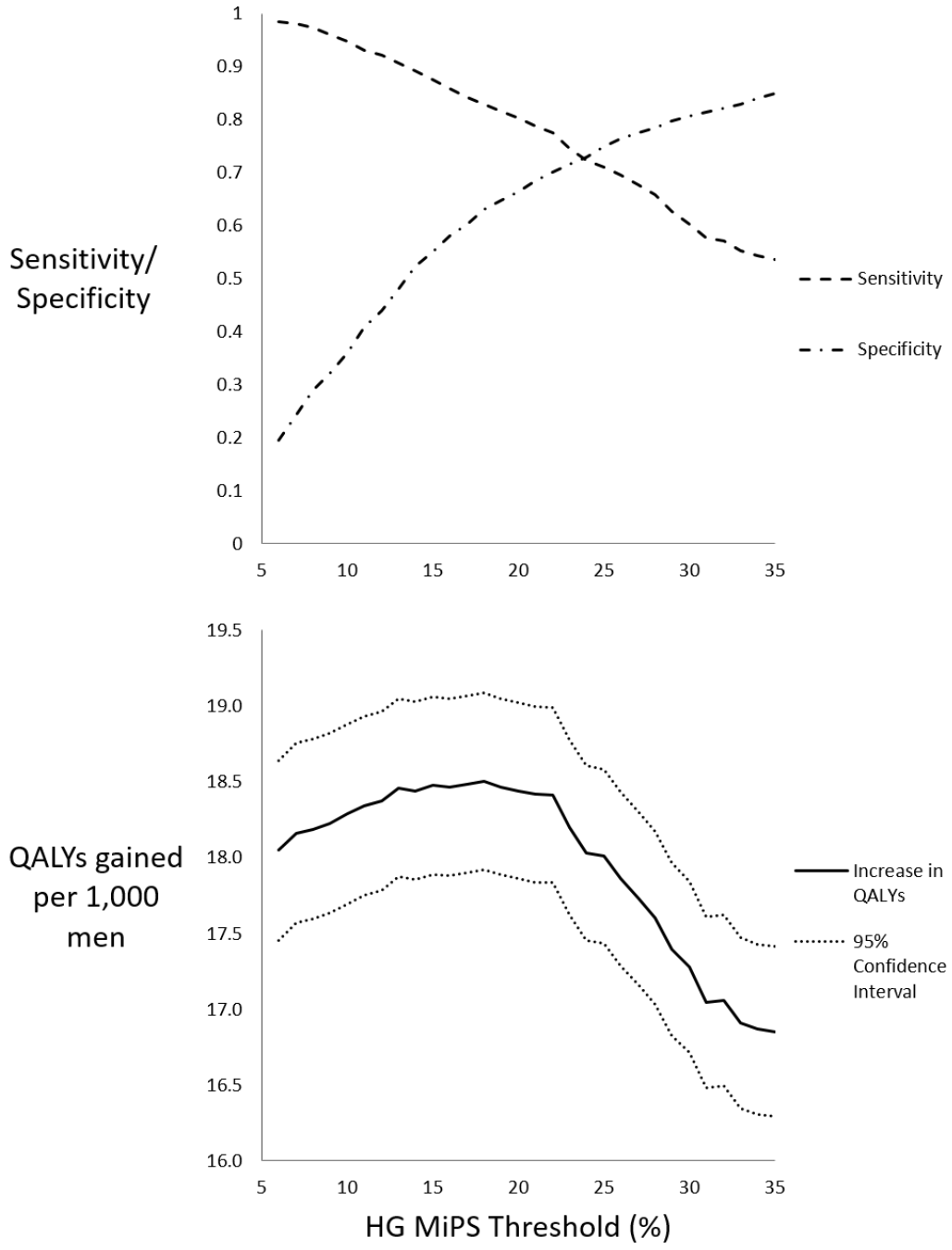

Figure 3.4: The relationship between sensitivity and specificity and their effect on the performance of the screening strategy for a range of high-grade (HG) MiPS thresholds to trigger biopsy. The performance of each threshold is assessed by calculating expected QALYs gained per 1000 men compared to no screening. Each of these strategies uses a PSA threshold of 2 ng/mL to trigger a high-grade MiPS test. The maximum QALY gain is achieved at a threshold of 18.

3.3.3 Sensitivity Analysis

We performed one-way and probabilistic sensitivity analysis on the screening strategy that maximized expected QALYs, which has a PSA threshold of 2 ng/mL and a second biomarker test with a high-grade sensitivity and specificity of 0.86 and 0.62, respectively. Using the base case parameter values, this 4Kscore strategy resulted in 27.7 prostate cancer deaths, 212 screening biopsies, and a gain of 19 QALYs per 1000 men. Additionally, we evaluated the impact of low screening adherence rates.

One-Way Sensitivity Analysis; One-way sensitivity analysis results are shown in Figure 3.5, which is a tornado diagram that displays the effect each parameter has on the expected increase in QALYs. The two parameters that had the greatest effect on expected gain in QALYs were d_t, the annual other-cause mortality rate, and δ_{Rec}, the annual QALY disutility for the 9-year post-radical prostatectomy recovery period.

Figure 3.6 presents the one-way sensitivity analysis on the number of prostate cancer deaths per 1000 men. The two parameters that had the greatest effect on the prostate cancer mortality rate were: d_t, the annual other-cause mortality rate, and w_t, the annual transition rate from No PCa to GS < 7 PCa, suggesting that patient groups that have a higher risk of developing prostate cancer (e.g. African Americans and patients with a family history) will be more likely to benefit from screening.

The parameter that had the greatest effect on both mean gain in QALYs and mean number of prostate cancer deaths was d_t, the annual other-cause mortality rate. When the low and high values of the annual other-cause mortality rate are used, the increase in QALYs per 1000 men ranged from 8 to 35 relative to the base case value of 19 QALYs, and the expected number of prostate cancer deaths per 1000 men ranged from 22.4 to 35.5 relative to the base case value of 27.7. This suggests that patients with comorbidities that are likely to have an increased risk of other-cause mortality may not receive as many benefits from screening.

Figure 3.5: One-way sensitivity analysis on expected gain in quality-adjusted life years (QALYs) per 1000 men relative to no screening. The model parameters that we varied are defined in Table 3.1.

Probabilistic Sensitivity Analysis; The probabilistic sensitivity analysis results are presented in Figure 3.7, which shows the number of screening biopsies versus the number of prostate cancer deaths per 1000 men from 30 experiments. The number of prostate cancer deaths ranged from 19.2 to 33.8, while the number of screening biopsies ranged from 196 to 215 per 1000 men.

Varying Adherence Rates; As mentioned previously, patient adherence to screening is often imperfect. To analyze the impact of adherence, we varied screening participation and attendance rates for the screening strategy that maximized expected QALYs, which has a PSA threshold of 2 ng/mL and a second biomarker test with a high-grade sensitivity and specificity of 0.86 and 0.62, respectively. We defined the participation rate as the proportion of patients that participate in screening. This is particularly relevant to prostate cancer screening, because fewer patients are

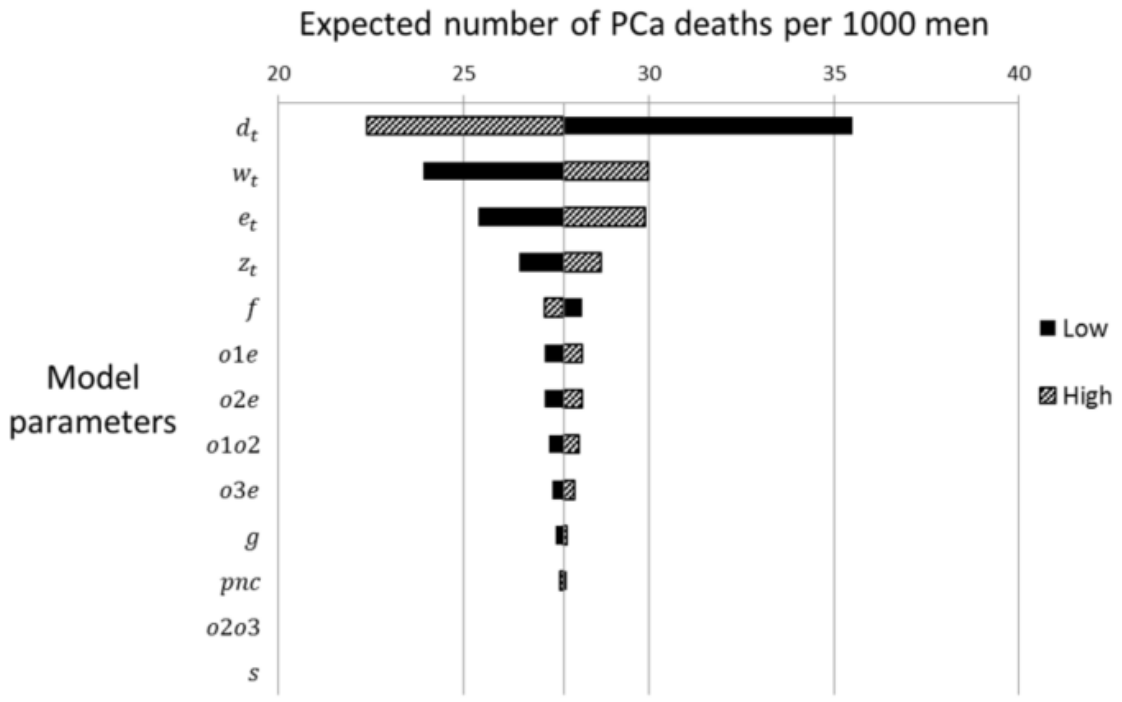

Figure 3.6: One-way sensitivity analysis on expected number of prostate cancer (PCa) deaths per 1000 men relative to no screening. The model parameters that we varied are defined in Table 3.1.

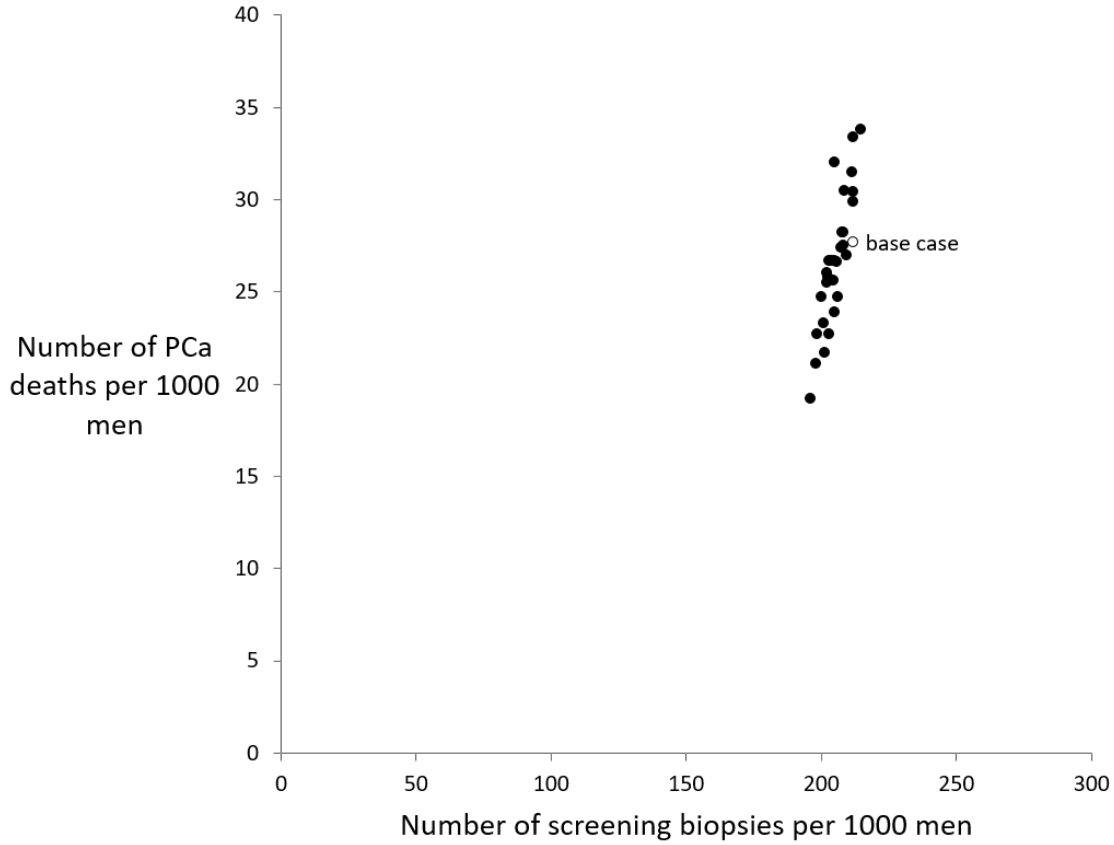

Figure 3.7: Probabilistic sensitivity analysis on the expected number of prostate cancer (PCa) deaths and screening biopsies per 1000 men. The model parameters that we varied and their bounds are defined in Table 3.1. The base case value on the figure is labeled, and the other points represent the 30 experiments.

Table 3.7: Strategy performance in terms of QALYS gained relative to no screening, number of screening biopsies, and number of prostate cancer (PCa) deaths per 1000 men after varying the screening participation rate in the population.

Screening participation rate	**Expected QALYs gained per 1000 men**	**Number of screening biopsies per 1000 men**	**Number of PCa deaths per 1000 men**
0.50	9.02	106.0	30.0
0.60	11.28	127.1	29.5
0.70	12.79	148.3	29.1
0.80	14.88	163.6	28.6
0.90	16.72	190.7	28.6
1.00	18.59	211.9	27.7

Table 3.8: Strategy performance in terms of QALYS gained relative to no screening, number of screening biopsies, and number of prostate cancer (PCa) deaths per 1000 men after varying the screening attendance rate in the population.

Screening attendance rate	**Expected QALYs gained per 1000 men**	**Number of screening biopsies per 1000 men**	**Number of PCa deaths per 1000 men**
0.50	15.94	163.5	28.6
0.60	16.67	177.2	28.3
0.70	17.22	188.3	28.1
0.80	17.97	197.5	28.0
0.90	18.18	205.3	27.8
1.00	18.59	211.9	27.7

participating in the screening process since the U.S. Preventive Services Task Force recommended against prostate cancer screening. We define the attendance rate as the probability that patients participate in screening in a particular year. We varied each of these values from 50% to 100%.

Table 3.7 presents our results from varying the screening participation rate (i.e. the proportion of the population that participates in screening), and Table 3.8 presents our results from varying the screening attendance rate (i.e. the probability a patient shows up for screening in a particular year). As suspected, as participation and attendance rates increase, the QALY gains increases, the number of prostate cancer deaths decrease, and the number of biopsies being performed increases. We found that participation rates have significantly more impact on patient outcomes than attendance rates.

3.4 Conclusions

We developed and validated a new partially observable Markov model that considers prostate cancer screening and treatment decisions for a cohort of men, starting at age 40, through to end of life. We used this model to examine alternative choices of two-stage biomarker-based screening strategies based on newly discovered biomarkers. The screening strategy with a PSA threshold of 2 ng/mL and a second biomarker with high-grade sensitivity and specificity of 0.86 and 0.62, respectively, increased the number of QALYs per 1000 men by 19 QALYs compared to no screening and by 7 QALYs compared to using the PSA test alone with a threshold of 4 ng/mL. Our model predicts one prostate cancer death averted per 200 men screened, assuming men were screened annually from age 50 to 75 with a PSA threshold of 4 ng/mL. *Gulati et al.* (2011) reported similar findings with a number needed to screen between 186 and 220.

Two recent modeling studies also examined the use of new biomarkers for prostate cancer screening. *Birnbaum et al.* (2015) and *Heijnsdijk et al.* (2016) evaluated the use of PCA3 and phi, respectively. We build on this previous work by evaluating many new biomarkers head-to-head in the same model, providing useful information when choosing between the many new biomarkers available. Another key difference from both of these studies is that we evaluated how the trade-off in sensitivity and specificity affects performance of new biomarkers, including hypothetical perfect biomarkers that provide an upper bound on the potential benefits of new biomarkers. Finally, we evaluated the biomarkers in the context of QALYs as well as prostate cancer deaths and number of biopsies per 1000 men.

A related study, *Merdan et al.* (2015), that considered the use of PCA3 and T2:ERG for repeat biopsy decisions found similarly significant reductions in the number of biopsies in this more specific context. We found that using an initial PSA threshold with a high sensitivity (2 ng/mL) and a second biomarker that has a high

sensitivity (between 0.68 and 0.95) and low to moderate specificity (between 0.36 and 0.78) to high-grade disease appears to maximize expected QALYs. Interestingly, high specificity in the second biomarker test, which is concomitant with low sensitivity, results in significant reduction in QALYs, but minimizes the number of screening biopsies. In our model, there are two populations of prostate cancer patients: (1) patients with low-grade disease (Gleason score 6), and (2) patients with high-grade disease (Gleason score ≥ 7). Patients with low-grade disease are unlikely to die from prostate cancer, and therefore, are unlikely to benefit from screening. Patients with high-grade cancer are more likely to develop metastatic disease, which is very likely to cause prostate cancer death. Thus, biomarker tests for high-grade cancer outperform all-cancer biomarkers for two reasons: (1) they are more likely to detect high-grade disease and prevent a prostate cancer death, and (2) these high-grade biomarkers reduce the number of biopsies for patients with low-grade disease reducing the burden of screening on patients that are unlikely to benefit.

In our one-way sensitivity analysis, we found that other-cause mortality has the greatest impact on the expected increase in QALYs relative to no screening, suggesting that the presence of comorbidity is an important consideration when determining the optimal prostate cancer screening strategy. We found that the results were most sensitive to variation in the QALY disutilities and the metastasis rate for patients with undiagnosed prostate cancer, and least sensitive to variation in transition probabilities. In our probabilistic sensitivity analysis, the prostate cancer mortality rate was more sensitive to variation in model parameters than the mean number of biopsies.

Many different screening strategies performed similarly in terms of QALYs; however, we have found that it is possible to distinguish these similar screening strategies by looking at additional performance measures that may better account for patient preferences. For example, some strategies that achieved similar QALYs varied significantly in rates of biopsy and prostate cancer deaths, with reductions in prostate

cancer deaths coming at the expense of a greater biopsy rate. This trade-off emphasizes the importance of a shared decision making approach to account for patient preferences regarding risk of prostate cancer mortality and harms from biopsy.

The hypothetical biomarkers that perfectly detect all cancer and high-grade cancer performed significantly better than screening strategies based on sensitivities and specificities reported in the literature. This suggests there may be potential for additional gains from new biomarker discoveries. Interestingly, the high-grade hypothetical perfect biomarker achieved similar rates of prostate cancer mortality when compared to the perfect all cancer biomarker, while reducing the number of screening biopsies patients are subjected to. These data suggest screening biomarkers with an ability to detect high-grade cancers may reduce unnecessary biopsies.

Our study has some limitations based on assumptions used in the modeling process. First, estimates of sensitivity and specificity for biomarkers can be dataset-dependent, as the estimates come from different datasets and, therefore, may have different biases; however, our analysis still provides useful insights into how the sensitivity and specificity of biomarkers impact long-term health outcomes. Second, we are not aware of any longitudinal studies of long-term health outcomes associated with these new biomarkers. In the absence of data to support correlations between disease status, risk of preclinical progression and recurrence, PSA levels, and new biomarkers operating characteristics, we have assumed no explicit correlations. If correlations exist, it could lead to biased results and conclusions. Lastly, we assumed that each patient receives at most one screening biopsy in his life. About $7 - 12\%$ of men undergoing biopsy have had a previous negative biopsy (*Nguyen et al.* (2010); *Thompson et al.* (2006)); however, the majority of patients receive a single biopsy, and cancers detected on second biopsy are typically less clinically significant. Since our intent is to measure the public health impact of biomarker screening, we do not believe this assumption significantly influenced our results.

These limitations notwithstanding, a number of conclusions can be drawn from this study. Identifying biomarkers and risk thresholds optimized for identification of high-grade cancers has the greatest impact on measures of performance in the screening setting. Combining new biomarkers with PSA has the potential to reduce the number of screening biopsies (thus decreasing overdiagnosis) and decrease the rate of prostate cancer mortality. The sensitivity analysis suggests our conclusions are robust with respect to plausible variation in model parameters. New biomarkers with risk thresholds optimized for identification of high-grade cancer can reduce the number of prostate cancer deaths compared to PSA alone, while also increasing quality-adjusted survival. These results support prospective clinical-validation trials using rationally selected thresholds in order to design more efficient strategies for the early detection of prostate cancer. We have shown that two-stage biomarker screening strategies can be beneficial for the early detection of prostate cancer and have provided a foundation for how this approach could potentially be adapted for other types of cancer screening.

CHAPTER IV

Cost Effectiveness of Magnetic Resonance (MR) Imaging and Targeted MR/Ultrasound Fusion Biopsy for Prostate Cancer Screening

4.1 Introduction

Concerns about the poor sensitivity and specificity of the PSA test have led to recommendations to discontinue prostate cancer screening in the United States (*Moyer* (2012)). In Chapter III we discuss the potential use of new molecular biomarkers in patients with elevated PSA to better select men for initial biopsy. MRI has recently been proposed as another potential minimally invasive way to achieve early detection of prostate cancer. MRI has higher sensitivity and specificity to high-grade disease than the new biomarkers we evaluated in Chapter III. Additionally, MRI could potentially reduce overtreatment by preferentially detecting intermediate- and high-grade cancers (*Siddiqui et al.* (2015); *Meng et al.* (2016); *Oberlin et al.* (2016); *Siddiqui et al.* (2016)); however, MRI is more costly than molecular biomarkers and there is limited evidence for its effectiveness as an intermediate test in patients being screened for prostate cancer. Moreover, there are multiple ways to use MRI in a screening setting, and it is not clear which is best. For example, if an MRI does not detect lesions suspicious for prostate cancer, either no biopsy or a standard biopsy

(which randomly samples cores of tissue from the entire prostate gland) can be performed. If an MRI detects suspicious lesions, a targeted MR/ultrasound fusion biopsy (i.e. targeted fusion biopsy) can be performed in which the MR images are used with real-time ultrasound to sample cores of tissue directly from suspicious lesions; alternatively, a combined approach can be used in which both standard and targeted fusion biopsies are performed during a single biopsy session. Since there are multiple ways to implement MRI in a screening setting, the optimal clinical pathway is unknown.

We used a Markov model to evaluate the cost-effectiveness of MRI in a screening setting. We used the model to predict outcomes for five screening strategies and report the results on the basis of 1000 men. The frequency of screening for each strategy was based on the AUA guideline for PSA screening (*Carter et al.* (2013)). The first strategy employed standard biopsy for men with elevated PSA (> 4 ng/mL). The other four strategies performed MRI on men with elevated PSA, and the results were used to decide whether the men should be referred for no biopsy, standard biopsy, targeted fusion biopsy, or combined (standard + targeted fusion) biopsy. We estimated the number of deaths averted, QALYs, and total cost for each strategy. Additionally, we estimated the incremental cost-effectiveness ratios (ICERs).

4.2 Model

We adapted the partially observable Markov model described in Chapter III to estimate outcomes for five screening strategies that utilize MRI. We also updated the annual metastasis rate based on the following estimates from the literature (*Johansson et al.* (2004)), with 95% confidence intervals shown in parentheses:

$$e_t = \begin{cases} 0.024\ (0.016 - 0.035), & \text{if } t \leq 70 \\ 0.015\ (0.009 - 0.026), & \text{if } t \geq 71 \end{cases}$$

Table 4.1: Definitions of five screening strategies.

Screening strategy	**PSA >4 ng/mL**	**Positive MRI**	**Negative MRI**
1	Standard Biopsy	-	-
2	MRI	Targeted Fusion Biopsy	Standard Biopsy
3	MRI	Targeted Fusion Biopsy	No Biopsy
4	MRI	Combined Biopsy	Standard Biopsy
5	MRI	Combined Biopsy	No Biopsy

We adopted this revised model because estimates for QALYs gained from PSA screening validate well relative to another recent cost-effectiveness study of PSA screening (*Heijnsdijk et al.* (2015)). For example, *Heijnsdijk et al.* (2015) reports that screening from ages 55 to 69 with two-year intervals and a PSA threshold of 3 ng/mL with a 3.5% discount rate results in 83 life-years gained and 61 QALYs gained per 1000 men. Under the same conditions, our model estimates 71 life-years gained and 59 QALYs gained per 1000 men.

For each strategy, for simulation purposes we used 30,000,000 samples of biopsy-naïve men who were screened every two years from age 55 to 69 according to the AUA guideline. In strategy 1, a standard biopsy was recommended for elevated PSA (>4 ng/mL). The decision-rule diagram for strategies 2 through 5 is shown in Figure 4.1. Each strategy recommended MRI for elevated PSA, while actions based on the MRI results depended on the strategy as defined in Table 4.1. Our model focuses on initial biopsy decisions; thus, the screening strategy terminates after the patient receives an initial biopsy or two negative MRIs; however, the patient continues to make state transitions in the absence of screening until all-other-cause mortality or clinical detection and subsequent prostate cancer mortality.

The model was comprised of discrete health states based on Gleason score, which are not directly observable, but can be inferred from PSA and/or MRI subject to published estimates of sensitivity and specificity. For standard biopsy, the results were randomly sampled as either positive or negative, assuming a sensitivity to any

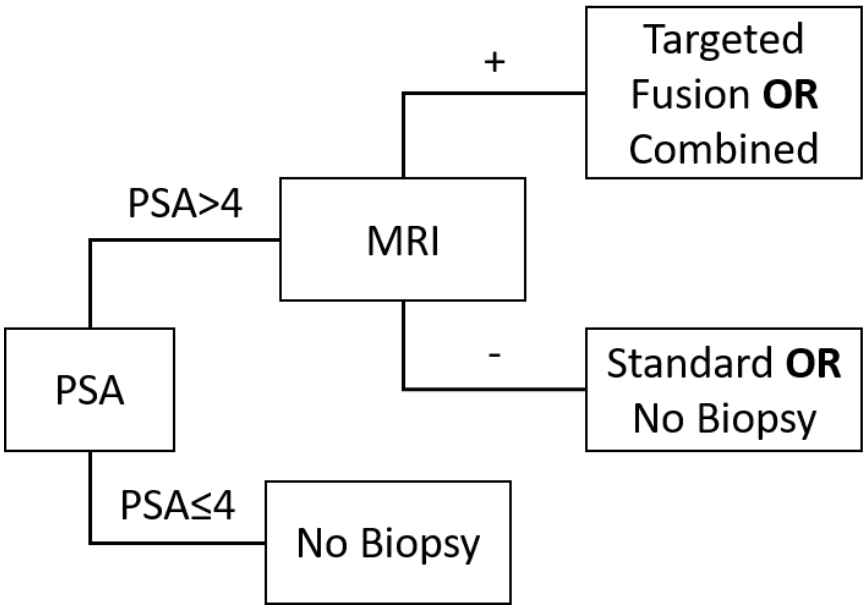

Figure 4.1: Decision rule diagram for screening strategies 2 through 5. All of the decision rules were compared to no screening and the case of standard biopsy for PSA greater than 4 ng/mL.

Table 4.2: Standard biopsy simulator based on data provided in *Epstein et al.* (2012).

		Biopsy result			
		No cancer	**GS <7**	**GS = 7**	**GS >7**
True health state	**No cancer**	1.00	0.00	0.00	0.00
	GS <7	0.20	0.71	0.08	0.00
	GS = 7	0.20	0.32	0.43	0.04
	GS >7	0.20	0.12	0.26	0.42

cancer of 0.8 (*Haas et al.* (2007)). If the biopsy result was positive, the probability that the biopsy provides an incorrect grading at diagnosis was based on data reported in *Epstein et al.* (2012). The exact standard biopsy data is provided in Table 4.2.

For targeted fusion and combined biopsy, we used the values of sensitivity and specificity to high-grade cancer reported in *Siddiqui et al.* (2015): 0.77 and 0.68, respectively, for targeted fusion biopsy, and 0.85 and 0.49, respectively, for combined biopsy. Based on Medicare infection rates reported in *Loeb et al.* (2011), 1.1% of biopsies performed led to hospitalization for post-biopsy infection (*Gonzalez et al.* (2012)).

As described in Chapter III, the model incorporated the clinical detection of symptomatic prostate cancer in addition to detection of prostate cancer through routine screening. For each patient, we randomly sampled a lead time from an elevated PSA measurement of ≥ 3 ng/mL to clinical diagnosis of prostate cancer from a distribution developed by *Savage et al.* (2010). After a patient had prostate cancer and a PSA score ≥ 3 ng/mL for their lead time and had not yet been diagnosed with prostate cancer, it was assumed the patient was clinically detected due to symptoms.

4.2.1 Treatment

In the updated model, patients with PSA > 20 ng/mL or a Gleason score ≥ 8 were assumed to receive a bone scan and a CT scan for staging (*Merdan et al.* (2014); *Risko et al.* (2014)). As described in Chapter III, patients with a biopsy result of Gleason score ≥ 7 received radical prostatectomy. Based on practice patterns reported in *Liu et al.* (2015), we assumed that 48.5% of patients diagnosed with Gleason score 6 prostate cancer received active surveillance, while the rest received radical prostatectomy. If a patient was clinically detected to have prostate cancer after age 80, we assumed they received watchful waiting. Men on active surveillance received an annual PSA test and a biopsy every two years. For men with no indication of progression, survival was consistent with survival for men with Gleason score 6 disease. If any biopsy indicated progression in Gleason score, the patient received radical prostatectomy. Men treated via radical prostatectomy had survival consistent with a treated population (*Roehl et al.* (2004)), with the potential for progression to metastatic prostate cancer and prostate cancer mortality. Other-cause mortality was based on estimates from CDC life tables (*Arias* (2010)).

Table 4.3: Clinical interpretation of PI-RADS scores (*Barentsz et al.* (2012)).

PI-RADS score	Clinical interpretation
1	Clinically significant disease is highly unlikely to be present
2	Clinically significant cancer is unlikely to be present
3	Clinically significant cancer is equivocal
4	Clinically significant cancer is likely to be present
5	Clinically significant cancer is highly likely to be present

4.2.2 PSA and MRI Sensitivity and Specificity

A published statistical model from the Prostate Cancer Prevention Trial was used to sample age-dependent and cancer onset-dependent PSA scores (*Gulati et al.* (2010)). The outcome of MRI was based on prostate imaging reporting and data system (PI-RADS) scores, between 1 and 5, which are defined in Table 4.3 with an increasing score indicating an increasing likelihood of the presence of clinically significant cancer (*Barentsz et al.* (2012)). We considered two PI-RADS thresholds to trigger biopsy: 3 and 4. A PI-RADS threshold of ≥ 3 had a sensitivity and specificity to clinically significant disease of 0.965 (95% CI: 0.868–0.994) and 0.597 (95% CI: 0.512–0.677), respectively, and a PI-RADS threshold of ≥ 4 had sensitivity and specificity values of 0.789 (95% CI: 0.658–0.882) and 0.789 (95% CI: 0.699–0.841), respectively (*Grey et al.* (2015)). Table 4.4 reports the probability that MRI results will be positive and negative for each threshold for a patient with no prostate cancer, prostate cancer with a Gleason score < 7, and prostate cancer with a Gleason score ≥ 7. To calculate the values in Table 4.4, we extracted the data from Figure 2 of *Grey et al.* (2015).

4.2.3 Costs and Quality of Life

For each combination of the five screening strategies and the two PI-RADS score thresholds, we estimated the mean cost and the mean QALYs gained per 1000 men relative to no screening. The values of the disutilities with their sources are shown

Table 4.4: The probability of positive and negative MRI results for different PI-RADS thresholds for no prostate cancer, Gleason score < 7 prostate cancer, and Gleason score ≥ 7 prostate cancer (*Grey et al.* (2015)).

	No prostate cancer		**Gleason score** < 7		**Gleason score** ≥ 7	
PI-RADS threshold	$\mathbb{P}(+$ MRI$)$	$\mathbb{P}(-$ MRI$)$	$\mathbb{P}(+$ MRI$)$	$\mathbb{P}(-$ MRI$)$	$\mathbb{P}(+$ MRI$)$	$\mathbb{P}(-$ MRI$)$
≥ 3	0.387	0.613	0.525	0.475	0.967	0.033
≥ 4	0.204	0.796	0.311	0.689	0.796	0.204

in Table 4.5. Our assumptions were similar to those of previous studies (*Aizer et al.* (2015); *Roth et al.* (2016); *Heijnsdijk et al.* (2012)). The post-recovery period for radical prostatectomy was assumed to last 9 years (*Heijnsdijk et al.* (2012)). *Li et al.* (2016) reported the disutility for hospitalization due to post-biopsy infection to be 0.28, which we assumed lasted for three weeks (*Heijnsdijk et al.* (2012)). *Grann et al.* (2011) reported the disutility for MRI as 0.04, which we assumed lasted for one week (*Heijnsdijk et al.* (2012)).

The reward update function for QALYs was:

$$\begin{aligned} q_t(s_t, a_t) = 1 - \Delta_{\text{Scr}}(a_t) - \Delta_{\text{MRI}}(a_t) - \Delta_{\text{Biop}}(a_t) - \Delta_{\text{Inf}}(a_t) - \Delta_{\text{Dia}}(a_t) \\ -\Delta_{\text{AS}}(a_t) - \Delta_{\text{RP}}(a_t) - \Delta_{\text{Rec}}(a_t) - \Delta_{\text{Met}}(s_t) - \Delta_{\text{Term}}(s_t) \end{aligned} \tag{4.1}$$

where $q_t(s_t, a_t)$ is the reward a patient receives at age t, which is 1 minus the disutilities associated with screening, MRI, biopsy, post-biopsy infection, diagnosis, active surveillance, radical prostatectomy, recovery from radical prostatectomy, metastasis, and terminal disease, as defined in Table 4.5. The arguments for the reward are the health state, s_t, that defines the cancer status of the patient and the action, a_t, that defines whether screening tests or biopsy was performed. Since we are also analyzing costs, we used discount factor of 3%. Thus, the number of discounted QALYs a patient receives at age t is:

$$\frac{q_t(s_t, a_t)}{(1.03)^{t-40}} \tag{4.2}$$

The total expected discounted QALYs a patient receives in their lifetime under screening policy i is:

$$Q^i = \mathbb{E}^i \left[\sum_{t=40}^{T} \frac{q_t(s_t, a_t)}{(1.03)^{t-40}} \right] \tag{4.3}$$

where T denotes maximum lifespan and the expectation is with respect to the stochastic process induced by the screening strategy i that defines the screening pathway and the thresholds at which to perform biopsies. Since exact evaluation of Q^i in equation 4.3 is not straight-forward due to history dependence of rewards up to a given decision epoch t, we used forward simulations to obtain statistical estimates:

$$\hat{Q}^i = \frac{1}{N} \sum_{n=1}^{N} \sum_{t=40}^{T} \frac{q_t^{(n)}(s_t, a_t)}{(1.03)^{t-40}} \tag{4.4}$$

Cost estimates with their sources are shown in Table 4.6. At each age, the cost of prostate cancer screening and treatment, c_t, is calculated. The discounted cost at age t with a discount rate of 3% is:

$$\frac{c_t}{(1.03)^{t-40}} \tag{4.5}$$

The total expected discounted cost in a patient's lifetime under screening policy i is:

$$C^i = \mathbb{E}^i \left[\sum_{t=40}^{T} \frac{c_t}{(1.03)^{t-40}} \right] \tag{4.6}$$

Since exact evaluation of C^i in equation 4.6 is not straight-forward due to history dependence of costs up to a given decision epoch t, we used forward simulations to obtain statistical estimates:

$$\hat{C}^i = \frac{1}{N} \sum_{n=1}^{N} \sum_{t=40}^{T} \frac{c_t^{(n)}}{(1.03)^{t-40}} \tag{4.7}$$

Table 4.5: Annual disutilities for health states considered in our cost-effectiveness analysis.

Health state	**Annual disutility (range)**	**Source**
PSA screning	0.00019 (0.0–0.00019)	*Heijnsdijk et al.* (2012)
MRI	0.00077 (0.00038–0.0012)	*Grann et al.* (2011)
		Heijnsdijk et al. (2012)
Biopsy	0.00577 (0.00346–0.0075)	*Heijnsdijk et al.* (2012)
Post-biopsy infection	0.0161 (0.00969–0.0291)	*Li et al.* (2016)
		Heijnsdijk et al. (2012)
Diagnosis	0.0167 (0.0125–0.0208)	*Heijnsdijk et al.* (2012)
Radical prostatectomy	0.247 (0.0917–0.323)	*Heijnsdijk et al.* (2012)
Post-radical prostatectomy recovery	0.05 (0.0–0.07)	*Heijnsdijk et al.* (2012)
Active surveillance	0.03 (0.0–0.15)	*Heijnsdijk et al.* (2012)
Palliative therapy	0.4 (0.14–0.76)	*Heijnsdijk et al.* (2012)
Terminal illness	0.3 (0.3–0.38)	*Heijnsdijk et al.* (2012)

Table 4.6: Costs considered in our cost-effectiveness analysis. Costs from the literature have been updated to 2016 US dollars based on inflation.

Intervention	**Unit costs in $**	**Source**
PSA screening	33.86	Medicare data
MRI	964.21	Medicare data
Standard prostate biopsy [a]	2,953.67	Medicare data
Targeted fusion prostate biopsy [b]	3,018.35	Medicare data
Combined prostate biopsy [b]	3,018.35	Medicare data
Post-biopsy infection-related hospitalization	6,361.31	*Adibi et al.* (2012)
		Gonzalez et al. (2012)
Staging	1,059.28	Medicare data
Active surveillance – standard biopsy (per year) [c]	1,642.58	Medicare data
Active surveillance – targeted biopsy (per year) [c]	1,674.92	Medicare data
Active surveillance – combined biopsy (per year) [c]	1,674.92	Medicare data
Radical prostatectomy	15,752.37	*Aizer et al.* (2015)
Distant-stage initial treatment	17,831.29	*Roth et al.* (2016)
Distant-stage management (per year)	2,500.65	*Roth et al.* (2016)
Other cause of death	5,975.15	*Mariotto et al.* (2011)
Prostate cancer death (age < 65)	103,884.24	*Mariotto et al.* (2011)
Prostate cancer death (age ≥ 65)	69,256.16	*Mariotto et al.* (2011)

[a] Includes professional, technical, and facility fees, pathology costs, and office visit.

[b] Includes professional, technical, and facility fees, pathology costs, office visit, and 3D reconstruction.

[c] Assumed to include an annual office visit, annual PSA test, and a biopsy every two years.

4.2.4 Cost Effectiveness

Future costs and QALYs were discounted to net present value using an annual discount rate of 3% (*Shepard* (1996)). Net costs per QALY gained were calculated for strategies 1 through 5 relative to no screening as the incremental costs of the screening strategy divided by the incremental QALYs of the screening strategy.

We identified the efficient strategies by removing dominated strategies (i.e., strategies that are more expensive and less effective than another strategy) as well as strategies ruled out by extended dominance (i.e., strategies that have higher ICERs than a more effective strategy) (*Shepard* (1996)). The ICERs of the efficient policies were calculated as the incremental costs divided by the incremental health gains compared to the next most effective strategy:

$$ICER = \frac{C^a - C^b}{Q^a - Q^b}, \text{ where } Q^a > Q^b. \tag{4.8}$$

If the ICER is under \$100,000/QALY, the screening strategy is considered cost-effective (*Neumann et al.* (2014)).

4.2.5 Sensitivity Analysis

To evaluate the robustness of our results, we performed one-way sensitivity analysis on the ICER for the optimal screening strategy. Ranges of the QALY disutilities appear in Table 4.5. Cost estimates and other-cause mortality rates (*Arias* (2010)) were varied by ±20%. The sensitivity and specificity of PI-RADS threshold 3 were varied using the 95% confidence intervals reported in *Grey et al.* (2015). The annual metastasis rate for patients with undiagnosed prostate cancer was varied within the 95% confidence interval reported in *Johansson et al.* (2004). Finally, we varied the annual prostate cancer incidence rate within the 95% confidence interval reported in *Haas et al.* (2007). Threshold analysis was also performed on the sensitivity and

specificity of MRI and combined biopsy under the optimal strategy. Base case values of the sensitivity and specificity of MRI were 0.965 and 0.597, respectively, and base case values of the sensitivity and specificity of combined biopsy were 0.850 and 0.490, respectively. During threshold analysis, we simultaneously reduced the sensitivity and specificity of MRI and combined biopsy until it was no longer cost-effective to use MRI for screening.

4.3 Results

4.3.1 Base Case Analysis

Table 4.7 presents the deaths averted, life-years and QALYs gained, the costs, and cost-effectiveness estimates for each screening strategy. The largest 95% confidence interval for QALY and cost per patient reflecting Monte Carlo statistical error was less than 1% of the corresponding sample-mean point estimate. The net discounted costs per QALY gained compared to no screening for each screening strategy was below \$100,000/QALY. Strategy 5 with a PI-RADS threshold of 3 maximized expected QALYs and number of prostate cancer death averted, and had the lowest net cost per QALY gained at \$33,953/QALY.

Figure 4.2 compares the QALYs gained per 1000 men under a PI-RADS threshold of 3 versus a PI-RADS threshold of 4. For each strategy, a PI-RADS threshold of 3 outperforms 4 in QALYs gained. Figure 4.3 shows the QALYs gained per 1000 men when using a targeted fusion biopsy versus a combined biopsy after a positive MRI. In each case, performing a combined biopsy after positive MRI resulted in additional QALY gains compared to performing a targeted fusion biopsy.

Figure 4.4 shows the incremental effectiveness in QALYs versus the incremental cost for each strategy relative to no screening. Dominated strategies were simultaneously more expensive and less effective than at least one other strategy. Interestingly,

Table 4.7: Predicted effects, costs, and cost-effectiveness for various screening strategies per 1000 men. Screening strategies are defined in Table 4.1.

Screening strategy	**PCa deaths averted** [a]	**Life-years gained** [a]	**QALYs gained** [a]	**Costs × $1000**	**Net costs per QALYs gained (3% discounted)** [a]
No screening	-	-	-	12,413	-
Strategy 1	4.7	58.7	47.8 (47.2–48.3)	12,964	39,381
Strategy 2, PI-RADS≥ 3	5.2	64.1	53.0 (52.4 –53.5)	13,050	40,019
Strategy 2, PI-RADS≥ 4	5.1	63.0	51.9 (51.3 –52.5)	13,064	41,415
Strategy 3, PI-RADS≥ 3	5.2	64.3	53.9 (53.3 –54.5)	13,034	37,218
Strategy 3, PI-RADS≥ 4	4.9	60.3	50.9 (50.3 –51.4)	13,038	38,059
Strategy 4, PI-RADS≥ 3	5.8	71.4	59.2 (58.6 –59.8)	13,021	36,138
Strategy 4, PI-RADS≥ 4	5.5	68.7	56.8 (56.2 –57.5)	13,041	37,725
Strategy 5, PI-RADS≥ 3	5.9	72.6	60.7 (60.1 –61.3)	13,002	33,953
Strategy 5, PI-RADS≥ 4	5.5	67.8	57.2 (56.6 –57.8)	13,009	34,426

Effects and costs are shown without discount. Cost-effectiveness is calculated at 3% discount rate for costs and QALYs. In 2016 US dollars. PCa = prostate cancer; QALY = quality-adjusted life year.
[a] Compared with no screening.

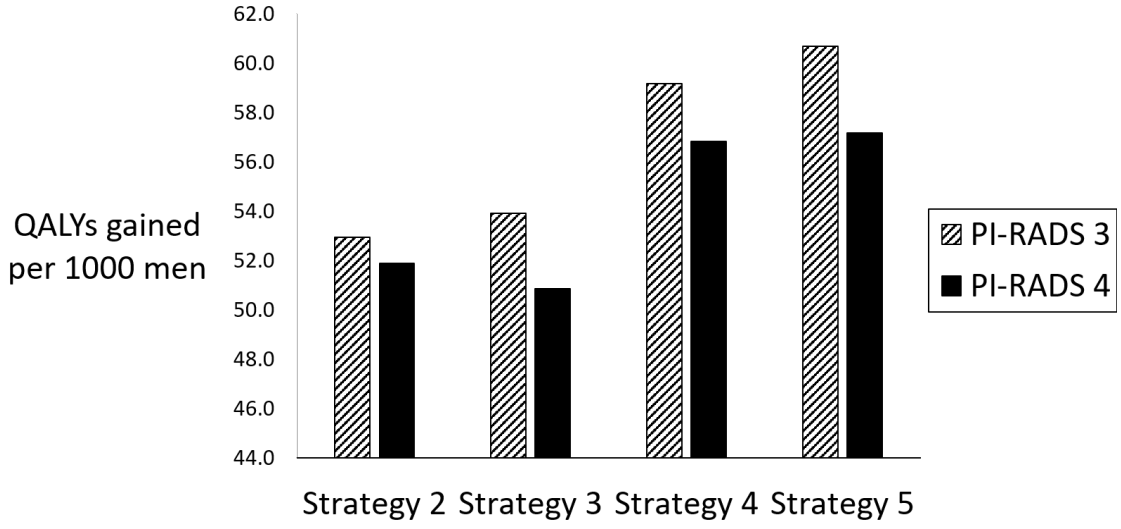

Figure 4.2: QALYs gained per 1000 men relative to no screening using a PI-RADS threshold of 3 versus 4 for Strategies 2–5. Strategy 1 resulted in 47.8 QALYs gained per 1000 men. Screening strategies are defined in Table 4.1. QALY = quality-adjusted life years; PI-RADS = prostate imaging reporting and data system.

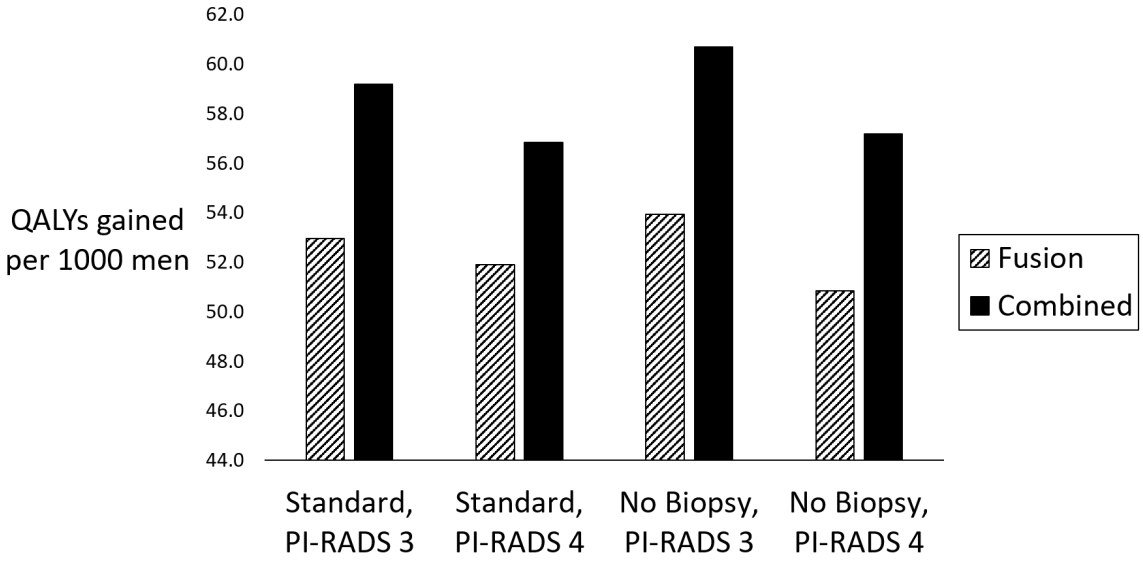

Figure 4.3: QALYs gained per 1000 men relative to no screening using a targeted fusion biopsy versus combined biopsy after positive MRI. Strategy 1 resulted in 47.8 QALYs gained per 1000 men. Columns are labeled with the type of biopsy performed after negative biopsy (no biopsy or standard biopsy) and the PI-RADS threshold used to indicate a positive MRI. QALY = quality-adjusted life years; PI-RADS = prostate imaging reporting and data system.

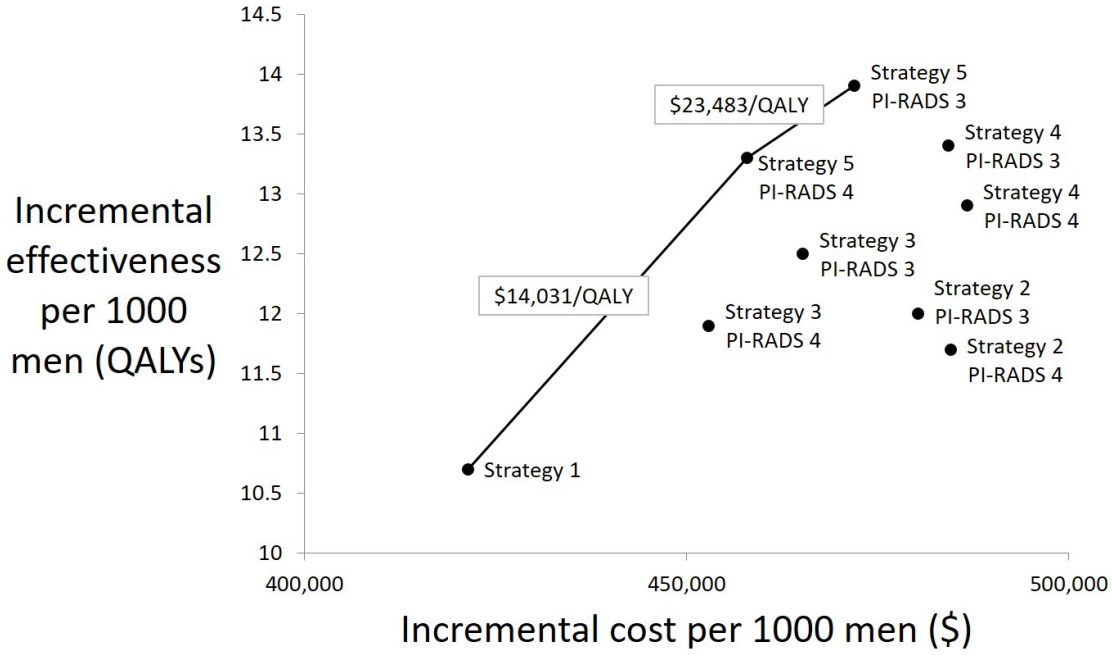

Figure 4.4: Incremental health benefits and costs associated with alternative screening strategies relative to no screening. Costs and QALYs are discounted at a rate of 3%. Each point is labeled with the screening strategy and PI-RADS threshold. Screening strategies are defined in Table 4.1. Lines connecting points representing two efficient screening strategies indicate the incremental cost-effectiveness ratio (ICER). QALY = quality-adjusted life years; PI-RADS = prostate imaging reporting and data system.

all four schemas that performed a standard biopsy after a negative MRI (strategies 2 and 4, with PI-RADS thresholds of 3 or 4) were dominated by strategies that performed no biopsy after negative MRI (strategies 3 and 5). The efficient strategies were strategy 1, strategy 5 with PI-RADS threshold of 4 with an ICER of $14,031/QALY, and strategy 5 with PI-RADS threshold of 3 with an ICER of $23,483/QALY. Thus, we found strategy 5 (i.e., MRI if PSA $>$ 4 ng/mL, combined biopsy if MRI positive, no biopsy if MRI negative) with PI-RADS threshold of 3 to be optimal under a willingness-to-pay threshold of $100,000/QALY.

4.3.2 Sensitivity Analysis

Figure 4.5 shows the one-way sensitivity analysis on the net costs per QALY gained relative to no screening for strategy 5 with a PI-RADS threshold of 3. We performed one-way sensitivity analysis on all model parameters; Figure 4.5 shows the parameters that varied the net costs per QALY gained by at least \$5,000/QALY when using the low and high values. The three model parameters that had the greatest impact were: (1) the metastasis rate for undiagnosed prostate cancer; (2) the annual QALY disutility for the 9-year post-radical prostatectomy recovery period; and (3) the annual QALY disutility for living with metastasis. In the sensitivity analysis, the only scenario that is not cost-effective under a willingness-to-pay threshold of \$100,000/QALY is a patient with a very low risk of developing metastasis, suggesting that our results are robust for most patients and cost-effective under a willingness-to-pay threshold of \$100,000/QALY. Threshold analysis shows that strategy 5 with a PI-RADS threshold of 3 remains cost-effective under a willingness-to-pay threshold of \$100,000/QALY when sensitivity and specificity of MRI and combined biopsy to high-grade cancer are all simultaneously reduced by 0.19. In particular, it is still cost-effective when sensitivity and specificity of MRI are ≥ 0.775 and ≥ 0.407, respectively, and sensitivity and specificity of combined biopsy are ≥ 0.660 and ≥ 0.300, respectively.

4.4 Conclusions

Based on our study, MRI as an intermediate test in the screening of men for prostate cancer is cost-effective assuming a willingness-to-pay threshold of \$100,000/QALY threshold. The optimal strategy was the use of MRI if PSA > 4 ng/mL, followed by combined biopsy if MRI was positive and no biopsy if MRI was negative, using a PI-RADS threshold of 3 to indicate a positive MRI. These results were robust

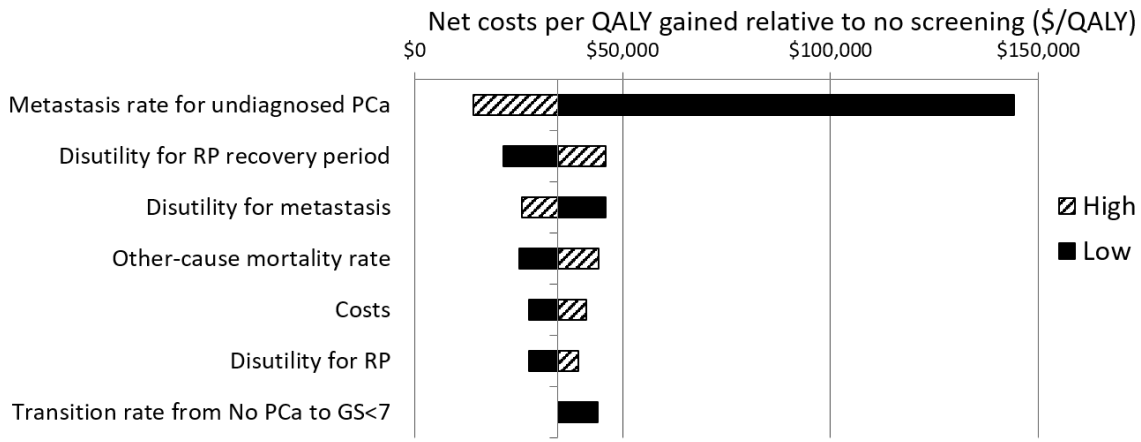

Figure 4.5: Tornado diagram of one-way sensitivity analysis on the net costs per QALY gained of strategy 5 with a PI-RADS threshold of 3 relative to no screening. Costs and QALYs are discounted at a rate of 3%. RP = radical prostatectomy; PCa = prostate cancer.

over a range of sensitivity analyses and were maintained even if the sensitivity and specificity of MRI and combined biopsy were reduced by 19 percentage points.

Although MRI has recently been proposed as an effective way to achieve early detection of prostate cancer, evidence in support of the use of MRI for early detection of prostate cancer in biopsy-naïve men is sparse. *Ahmed et al.* (2017) showed that MRI could be effective from a clinical perspective by reducing primary biopsy and clinically insignificant cancer diagnoses, but did not consider the cost-effectiveness. *Willis et al.* (2014) performed clinical decision analysis and *de Rooij et al.* (2014) performed cost-effectiveness analysis; however, both studies assumed a fixed sensitivity and specificity of MRI and assumed that positive MRI is automatically followed by a targeted fusion biopsy, while negative MRI automatically results in no biopsy. Thus, they evaluated one clinical pathway compared to the standard of care. Our study evaluated strategies that performed targeted fusion biopsy or combined biopsy on positive MRI, as well as the option to perform a standard biopsy or no biopsy on negative MRI. Thus, our study evaluated eight MRI-based clinical pathways (two PI-RADS thresholds for each

of the four MRI-based strategies) compared to screening with PSA alone, allowing us to estimate the effects of varying PI-RADS thresholds and biopsy techniques on the cost-effectiveness of using MRI for prostate cancer screening.

Additionally, our study focuses on long-term costs and health outcomes over the patient's entire lifetime, rather than assessing short-term outcomes. Including long-term costs and health impacts enabled us to assess the potentially negative impact of detecting low-risk cancers related to harm from biopsy(-ies) and overtreatment. Prior studies did not account for the costs and harms associated with biopsy complications, resulting in an overestimation of the benefit from screening and an underestimation of the costs.

Heijnsdijk et al. (2015) evaluated the cost-effectiveness of several PSA screening policies in the absence of MRI, and our models produced similar expected outcomes for PSA screening. The net cost per QALY gained we present for PSA screening is lower than the results reported in *Heijnsdijk et al.* (2015) because we include more costs in our model, including the significant cost of a prostate cancer-related death.

Using MRI for prostate cancer screening resulted in health benefits for the patient compared to both no screening and screening using PSA alone. For example, the screening strategy where men with a PI-RADS score ≥ 3 were recommended for combined biopsy (i.e., strategy 5) resulted in 5.9 prostate cancer deaths averted, 60.7 QALYs gained, and 72.6 life-years gained per 1000 men compared to no screening. For every screening strategy, a PI-RADS threshold of 3 outperformed a threshold of 4 in terms of QALYs, while also resulting in lower costs. Our results also suggest that performing a combined biopsy after a positive MRI outperforms performing a targeted fusion biopsy in terms of QALYs. However, there does not appear to be a benefit to performing standard biopsy on negative MRI, because it results in additional costs and disutility to the patient and overall does not provide sufficient health benefits. This conclusion has been supported in the literature. For example, *Hansen et al.*

(2016) concluded that biopsies may not be necessary for men with elevated PSA and nonsuspicious MRI because the negative predictive value for excluding Gleason score ≥ 7 disease on MRI was very high. Our study adds additional evidence in support of this conjecture.

Given the wide variability in the quality of radiology reporting and interpretation of MRI results, we performed threshold analysis on the sensitivity and specificity of MRI and combined biopsy. These analyses found this approach to be a cost-effective method of early detection even if the sensitivity and specificity were substantially lower than estimates reported in the literature, suggesting that our results may be relevant in a community setting where sensitivity and specificity may be lower than specialized medical centers where most previous MRI studies have been conducted. Sensitivity analysis suggests our results are robust with respect to reasonable variation of the model parameters; however, the results are sensitive to the annual metastasis rate for an undiagnosed prostate cancer patient. Under the metastasis rate assumption from Chapter III, MRI is not cost-effective when QALYs are discounted.

One potential limitation of our study is that there is the potential for bias in the data we used to estimate MRI results because the population used includes patients with previous negative biopsies in addition to biopsy-naïve patients; however, by using the estimates based on the larger patient population we were able to obtain better estimates of sensitivity and specificity. Our sensitivity analysis further confirms our conclusions are not sensitive to this assumption. Another possible limitation is the inconsistent definition of clinically significant prostate cancer in the literature. For example, *Siddiqui et al.* (2015) defined clinically significant disease as high-volume Gleason 3+4, or Gleason $\geq 4+3$, while *Grey et al.* (2015) defined clinically significant disease to be cancer core involvement ≥ 6 mm or the presence of any Gleason pattern 4. In our model, we considered clinically significant disease to be any Gleason score ≥ 7. Additionally, the only curative treatment included in our model was rad-

ical prostatectomy, because it is the most common curative treatment, and patients undergoing radiation therapy have similar health outcomes (*Hamdy et al.* (2016)). Finally, our model uses many different sources of data; however, given the long-term evaluation period needed for prostate cancer screening, randomized trials are unlikely to be able to assess long-term QALYs and costs. These limitations notwithstanding, we believe this study provides important evidence in support of the use of MRI for early detection of prostate cancer in biopsy-naïve men, both from a health benefit and cost perspective.

Our results show that incorporating MRI into prostate cancer screening in biopsy-naïve men is cost-effective under a willingness-to-pay threshold of \$100,000/QALY. The strategies that performed a standard biopsy on negative MRI were more expensive and less effective than strategies that perform no biopsy on negative MRI. The screening strategy where men with PI-RADS score ≥ 3 were recommended for combined biopsy, while men with PI-RADS score < 3 were recommended for no biopsy was optimal and cost-effective with an ICER of \$23,483/QALY. Therefore MRI appears to be a viable approach for early detection of prostate cancer from a cost-effectiveness perspective. More analysis would need to be done to explore whether the use of molecular biomarkers could also be cost-effective, since biomarkers are less expensive and have lower sensitivity and specificity to high-grade prostate cancer compared to MRI.

CHAPTER V

Optimization of Biomarker-Based Screening Policies

5.1 Introduction

The policies evaluated in Chapter III are myopic in the sense that they make biopsy decisions based solely on a patient's most recent test results, without considering the patient's full medical history. However, benign conditions can cause a sudden spike in a patient's biomarker scores, which motivates the potential to use Bayesian updating to estimate the *belief state* for patients so decisions can be made based on estimates of patient risk of cancer, rather than biomarker scores. Thus, we have extended the model of Chapter III to create a new POMDP model to investigate optimal prostate cancer screening decisions based on a patient's *belief state*, which is calculated using Bayesian updating and comprises a patient's complete history of biomarker test results in a way that is similar to the model first proposed by *Smallwood and Sondik* (1973). This POMDP can be used to determine how, if at all, new biomarker tests should be used for prostate cancer screening. We present results for the case of high-grade MiPS. We chose high grade MiPS because it was found to be a good biomarker in Chapter III and because we had access to the data necessary to estimate the probability distribution of MiPS conditional on the patient cancer

status. However, the approach we lay out could be applied to other biomarkers and in other disease contexts.

5.2 Model

We have developed a POMDP model that maximizes total expected QALYs by optimizing the decision to conduct a biopsy based on the patient's belief state at annual decision epochs. In this chapter, to be consistent with the literature on POMDPs we will refer to the patient's underlying health state as the *core state* of the patient. The set of core states and the one-step transition probability matrix are the same as defined in Chapter III: $S = \{\text{NC, OCG1, OCG2, OCG3, EPLN, PRFT, NRFT, M, D}\}$.

At each decision epoch from ages $t = 1, \ldots, T$, we assume a high-grade MiPS test is performed. The set of actions is $A = \{\text{Wait, Biopsy}\}$, i.e. wait until the next decision epoch or perform a biopsy. The observations that result from the high-grade MiPS test inform the action. The observation space for the high-grade MiPS test are continuous values between 0 and 1; however, to simplify the problem we discretized these observations into clinically relevant bins. The set of observations is Θ, which includes the MiPS discretized observations in addition to Post-treatment (PT), Metastasis (M), Death (D). The observations for the action "Biopsy" are NC, OCG1, OCG2, and OCG3. Since only a small amount of tissue is sampled during a biopsy, sampling error can result in a false negative or incorrect grading at diagnosis. Based on the discretization of the MiPS test results, we have developed information matrices by age, $Q_t, t = 55, \ldots, 69$. The information matrix has rows associated with the core health states and columns associated with the set of possible observations. We denote the components of each information matrix by $q_t(\theta|s_t)$, which defines the probability of observing $\theta \in \Theta$ at age t given the core state of the patient is $s_t \in S$.

In the following description of the model we use notation consistent with the notation used in *Smallwood and Sondik* (1973). Let $\pi^t = [\pi_1^t, \pi_2^t, \ldots, \pi_9^t]$ be the belief

vector, where π_i^t is the probability that the patient is in state i at decision epoch t and $\sum_i \pi_i^t = 1$. The belief vector is updated via Bayesian updating after the observations at each decision epoch, using the following equation:

$$\pi_j^{t+1} = \frac{\sum_i \pi_i^t P_t^{Tr}(i,j) q(\theta|j)}{\sum_{i,j} \pi_i^t P_t^{Tr}(i,j) q(\theta|j)},$$

where $q(\theta|j)$ denotes the probability of observing θ, given the core state of the patient is j. The numerator calculates the probability of transitioning to state j and observing output θ, and the denominator is the probability of observing output θ. This is an application of Bayes law, and the equation is developed in detail in Appendix A of *Smallwood and Sondik* (1973).

Our model seeks to maximize expected QALYs, using the following set of optimality equations:

$$V_t(\pi^t) = \max_{a_t \in A} \left\{ r(\pi^t, a, \pi^{t+1}) + \sum_{i,j,\theta} \pi_i^t P_t^{Tr}(i,j) q_t(\theta|j) V_{t+1}(j) \right\}, \quad t = 1, ..., T-1$$

$$V_T(\pi^T) = R(\pi^T)$$

where $r(\pi^t, a, \pi^{t+1})$ is the immediate reward in QALYs and $V_t(j)$ is the maximum expected future QALYs for a patient in state j at age t.

5.3 Methods

The terminal reward vector, α^T, consists of the expected QALYs for a patient at age 70 in each core state. The infinite state space of a POMDP makes it difficult to solve. Thus, we have divided the space into a fixed-finite grid, which allows us to approximate the infinite belief space, $B = \{\pi^t | \sum_i \pi_i^t = 1\}$, with a finite grid of points. By discretizing the belief space, we can associate every sampled belief state with one of the grid points (e.g. based on the closest grid point). Thus, the continuously

sampled states are mapped to a finite set of states.

At any given decision epoch, we know that the patient will be in one of three subsets of the state space: $S_1 = \{\text{NC, OCG1, OCG2, OCG3, EPLN}\}$, $S_2 = \{\text{PRFT, NRFT}\}$, or $S_3 = \{\text{M, D}\}$. The set S_1 includes the unobservable pre-diagnosis states, S_2 includes the unobservable post-treatment states, and S_3 includes the observable states. In other words, if a patient has not yet been diagnosed with prostate cancer, we know that they are in one of the five states in S_1; if a patient has been treated for prostate cancer, we know they are in one of the post-treatment states in S_2; and we know the exact state of the patient when they are in the completely observable S_3. Therefore, any grid point in the discretized state space we generate will have non-zero entries in only one of the three subsets. The grid points in S_3 will consist of $(0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 0)$ and $(0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1)$, i.e., the states in S_3 are perfectly observable.

We utilized a data-driven sampling approach to develop the grid at each age. We denote a sample path to be the stochastic progression of a patient's health states, biomarker test results, biopsy results, and belief states. A patient's sample path depends on the screening strategy being used. For this reason, we generated sample paths using varying policies. The sample paths provided information about not only where patients' beliefs are located in the belief space at each age, but also information about where a patient's belief's are not located in the belief space. Due to the computational burden of having a large number of grid points we then used these samples to create a smaller grid of points at each age using k-means clustering. The goal of k-means clustering is to divide L points (i.e., the sample belief states) in M dimensions (i.e., the partially observable health states) into k clusters so that within-cluster sum of squares is minimized (*Hartigan and Wong* (1979)). The centroid of each cluster then defines a unique grid point. Algorithm 1 gives a description of the k-means clustering algorithm and an example is given in Figure 5.1.

Algorithm 1 k-means clustering algorithm for grid development.

Inputs: x sample paths through the belief space; number of grid points at each age, k

for $t = 55, \ldots 69$ **do**
    Randomly divide the x sample belief states into k clusters.
    Calculate the centroid of each cluster and let that be the mean.
    while convergence is not reached **do**
        Cluster each simulated sample belief state with the nearest mean.
        Calculate the centroid of each cluster and let that be the new mean.
    end while
    The k means represent the k grid points for age t.
end for

Smallwood and Sondik (1973) presented the first exact method for POMDPs, known as the "One-Pass Algorithm". *Smallwood and Sondik* (1973) shows that the value function of a POMDP is piecewise linear and convex, and can be written as

$$V_t(\pi^t) = \max_k \left\{ \sum_{i=1}^{N} \alpha_i^k(t)\pi_i^t \right\},$$

for some set of vectors $\alpha^k(t) = [\alpha_1^k(t), \alpha_2^k(t), \ldots, \alpha_N^k(t)],\ k = 1, 2 \ldots$, which are referred to as α-vectors. Each α-vector has an action associated with it; therefore, we can evaluate each α-vector at a fixed belief state to determine the α-vector that maximizes the value function and the optimal action associated with it. However, there may be many α-vectors in the set that are not needed to define the value function (i.e. they are not a maximum α-vector over the entire belief space). The algorithm described in *Smallwood and Sondik* (1973) defines regions for an α-vector and searches for a belief where that α-vector is not dominant. The "One-Pass Algorithm" solves a series of linear programs to try to find a minimal α-vector set. However, we must solve a linear program for each α-vector in the set, which is computationally expensive since the set of α-vectors grows exponentially in the size of the set of possible observations.

Cassandra et al. (1994) proposed the "Witness Algorithm", which also defines

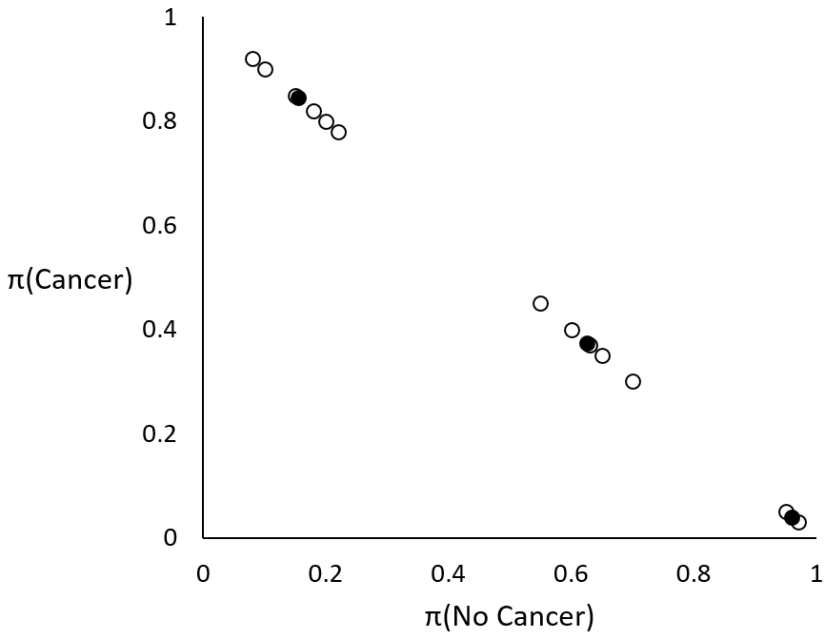

Figure 5.1: An example of k-means for a simple 2-dimensional case, where the unfilled points represent the $L = 13$ sample points and the filled points represent the $k = 3$ grid points.

regions for a vector and looks for a point where that vector is not dominant. As a result, the "Witness Algorithm" only needs to solve one linear program for each α-vector in the minimal α-vector set, which is an improvement on the "One-Pass Algorithm".

In our approach, we draw on the basic idea of the witness algorithm to accelerate the one-pass algorithm. Terminal rewards define the α-vector in decision epoch $T+1$. At each age moving backwards starting from decision epoch T, we use Monahan's algorithm (*Monahan* (1982)) to calculate the set of α-vectors in decision epoch t using the α-vectors from decision epoch $t+1$. We then eliminate any α-vectors that do not define the value function at one of the grid points that approximates the belief space, i.e., each grid point acts as a witness point for one α-vector and thus the number of α-vectors is limited to the number of grid points. We use our simulation model and the k-means clustering model to develop a set of grid points that represent the areas

of the belief space that our patient population is most likely to be, and use those grid points to calculate a subset of the *relevant* α-vectors by finding the dominant α-vector associated with each grid point. After we complete this step, we have developed a policy, where we determine the (approximated) optimal action at any point in the belief space by selecting the action associated with the α-vector that maximizes the value function, $\alpha^T \pi$.

5.4 Results

To develop age-dependent grids, we generated 600 randomly sampled patient sample paths through the belief space for 110 different MiPS policies. The 110 policies included every combination of the previously recommended schedules described in Table 5.1 with the following thresholds: {0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50}. We then used these samples to create a grid of points at each age using the k-means clustering heuristic in Matlab, where $k = 200$, $L = 66,000$, and $M = 5$. As shown in Figure 5.2, we found that at older ages there are more grid points with a higher belief of having prostate cancer, which agrees with the fact that the prevalence of prostate cancer increases at older ages.

Patients were screened at each decision epoch from ages 55 to 69. In our experiments, we discretized the continuous observations space of MiPS scores into three observations: $\Theta = \{[0, .125), [.125, .375), > .375\}$, which equally divided patients into low, median, and high risk groups. We evaluated two types of POMDP screening policies. The first was based on the policy developed using the data-driven sampling method above to prune α-vectors and create an approximation of the optimal policy subject to error induced by using a finite set of grid points. The second uses a risk-based threshold to trigger prostate biopsy based on a patient's current belief of having Gleason score > 7 or extraprostatic or lymph node-positive disease (i.e., the belief the patient is in states OCG3 or EPLN). Under this risk-based POMDP policy

Table 5.1: Screening schedules for the prostate cancer screening policies used to generate sample paths. The screening schedule defines the set of decision epochs during which screening occurs.

Schedule Label	Range of Ages (yr)	Screening Interval (yr)	Source
S1	40-75	5	*Ross et al.* (2000)
S2	50-75	2	*Ross et al.* (2000)
S3	50-75	1	*Ross et al.* (2000), *Andriole et al.* (2009)
S4	40,45	-	*Ross et al.* (2000)
	50-75	2	
S5	40,45	-	*Ross et al.* (2000)
	50-75	1	
S6	55-69	1	*Heijnsdijk et al.* (2012)
S7	55-74	1	*Heijnsdijk et al.* (2012)
S8	55-69	4	*Heijnsdijk et al.* (2012)
S9	55	-	*Heijnsdijk et al.* (2012)
S10	60	-	*Heijnsdijk et al.* (2012)
S11	65	-	*Heijnsdijk et al.* (2012)

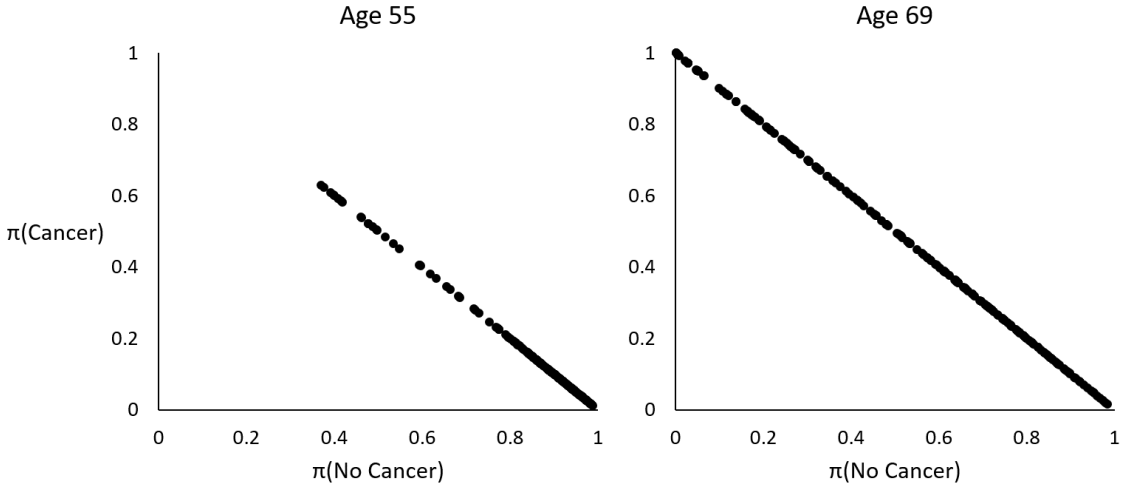

Figure 5.2: The grids generated for ages 55 and 69, where the belief of cancer is calculated by adding the belief of each of the four cancer states.

with a risk threshold of x, when $\pi_{\text{OCG3}} + \pi_{\text{EPLN}} > x$, the patient receives a biopsy. We found that a risk threshold of 0.35 maximized expected QALYs for the patient. We evaluated the resulting POMDP policies using our simulation model. We compared the results to the following six myopic screening policies based on commonly used thresholds:

- No screening
- PSA with a threshold of 4 ng/mL
- PSA with a threshold of 2 ng/mL to trigger a PCA3 test with a threshold of 25 (patients with a PSA $>$ 10 ng/mL will automatically receive a biopsy)
- PSA with a threshold of 2 ng/mL to trigger a T2:ERG test with a threshold of 10 (patients with a PSA $>$ 10 ng/mL will automatically receive a biopsy)
- High-grade MiPS test with a threshold of 15
- MiPS test with a threshold of 25

A PSA threshold of 4 is common, the two-stage policies were evaluated in Chapter III, and the MiPS and high-grade MiPS thresholds performed well in previous experiments.

We estimated the terminal reward vector using our simulation model of Chapter III with 10,000,000 sample paths starting at age 70 from each core state:

$$\alpha^T = \begin{array}{c} \begin{array}{ccccccccc} \text{NC} & \text{OCG1} & \text{OCG2} & \text{OCG3} & \text{EPLN} & \text{NRFT} & \text{PRFT} & \text{M} & \text{D} \end{array} \\ \left(\begin{array}{ccccccccc} 14.062 & 13.887 & 13.609 & 13.532 & 9.826 & 14.104 & 13.592 & 2.823 & 0 \end{array} \right). \end{array}$$

Each element of α^T is an estimate of expected remaining lifespan for the corresponding state.

Figure 5.3 shows the results in terms of QALYs with confidence intervals for 10,000,000 samples. We found that by using the policy generated by our approximated POMDP solution which accounts for a patient's entire history of their biomarker test results, we can gain 181.7 QALYs, which significantly outperformed all myopic policies. Thus, we found that it is possible to develop screening policies using our approximated solution technique on a discretized POMDP, and that it results in health benefits for the patient. The POMDP policy with a risk threshold of 0.35 gained 193.4 QALYs per 1000 men. The difference between these POMDP policies was not statistically significant. The POMDP approximated solution depends on the grid selection, while the risk-based POMDP policy does not, suggesting there is potential of increased gains through better grid selection.

5.5 Conclusions

In this chapter, we presented a new POMDP model to estimate optimal biopsy screening policies based on a patient's *belief state* rather than their latest biomarker test results. The underlying health states of the patient are not directly observable; however, their high-grade MiPS results provide some information about their core health state. Patients were screened annually with the high-grade MiPS test from ages 55 to 69. Due to the large number of observations and unobservable states, we presented a data-driven approximation method to solve this POMDP. We found that it is possible to develop screening policies using our approximated solution technique on a discretized POMDP, and that it results in significant health benefits for the patient. We also found that an easier to implement risk threshold based policy has similar performance to the optimal solution to the discretized POMDP.

Our study has limitations based on assumptions used in the modeling process. First, active surveillance and radical prostatectomy were assumed to be the only treatment options, because radical prostatectomy is the most common curative treatment,

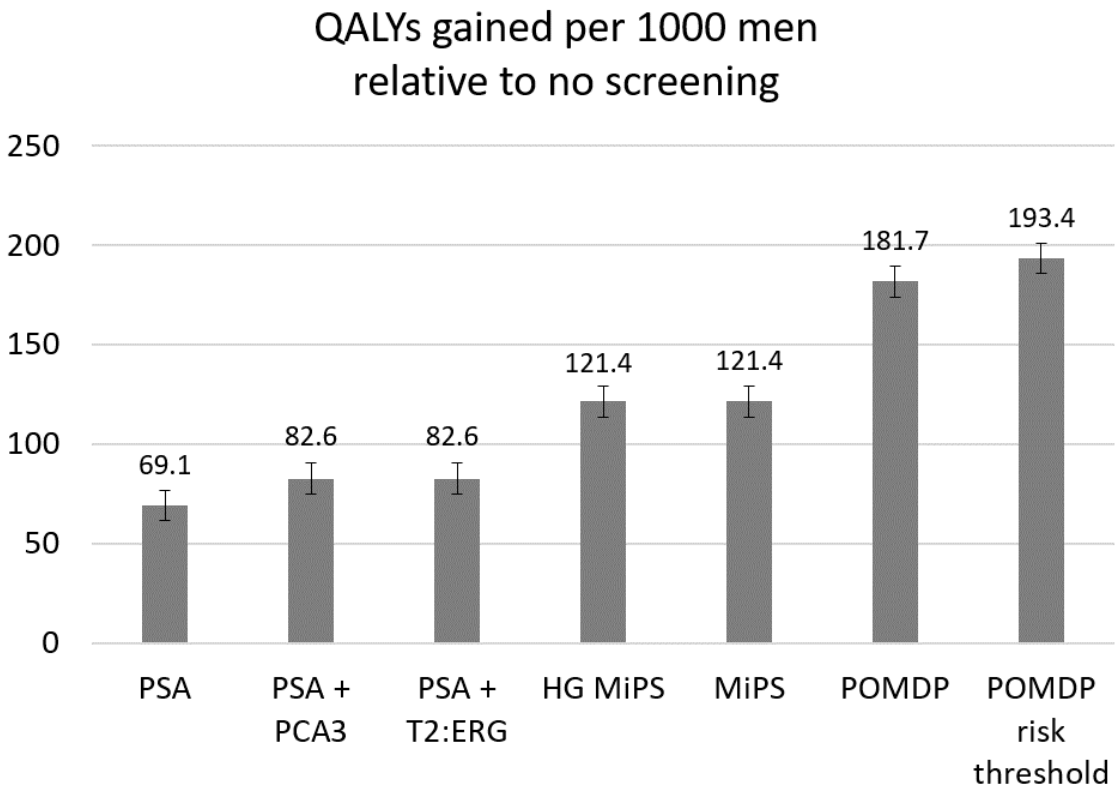

Figure 5.3: Expected increase in QALYs per 1000 men compared to no screening for a range of myopic policies compared to two policies based on our POMDP with three HG MiPS observations. The first POMDP policy was generated by our approximated POMDP solution and the second POMDP policy performs a biopsy when a patient's belief of having OCG3 or EPLN is ≥ 0.36.

and patients undergoing radiation therapy have similar health outcomes (*Hamdy et al.* (2016)). Lastly, we assumed that each patient receives at most one screening biopsy in his life. About $7 - 12\%$ of men undergoing biopsy have had a previous negative biopsy (*Nguyen et al.* (2010); *Thompson et al.* (2006)); however, the majority of patients receive a single biopsy, and cancers detected on second biopsy are typically less clinically significant. Since our intent is to measure the public health impact of biomarker screening, we do not believe this assumption significantly influenced our results.

5.6 Future Work

Both POMDP policies depended on the discretization of the continuous observation space. In our experiments, we discretized the continuous observation space into three observations. In the future, we will explore whether expanding the number of observations in the discrete approximation has an impact on long-term health outcomes. Additionally, our experiments had approximately 200 grid points for each age. The POMDP approximated solution depends on the grid selection, while the risk-based POMDP policy does not. The risk-based POMDP policy outperforming the approximated POMDP solution which suggests there could be potential additional health gains that could be achieved through better grid selection. We propose two different ways to improve the grid. First, we could simply increase the number of grid points at each age. Second, we could develop a closed-loop algorithm, where the policy developed by our POMDP approximation technique informs our grid point selection. A diagram showing how this closed-loop would work is shown in Figure 5.4. By improving our discretization of the continuous observation space and belief space, we can observe what impact these parameters have on the policy and the resulting long-term health outcomes for the patient.

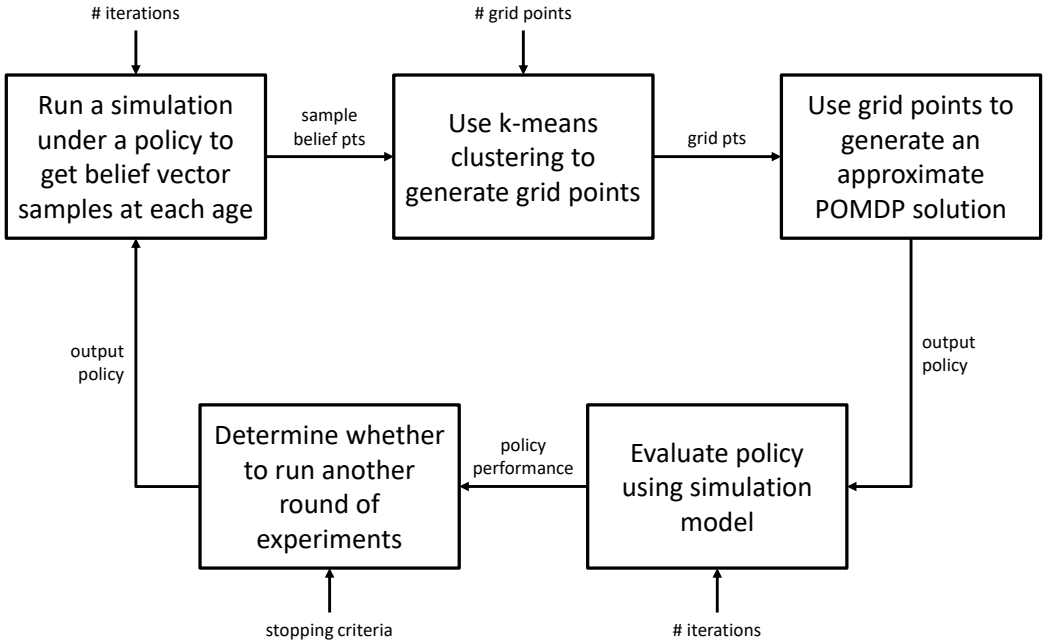

Figure 5.4: Closed-loop diagram showing how we will use the policy generated by our POMDP approximation technique to develop a new more-relevant grid. Arrows pointing to a box indicate inputs that are needed in the process.

CHAPTER VI

A Hidden Markov Model for Optimizing Active Surveillance Strategies for Low Risk Prostate Cancer

6.1 Introduction

Although prostate cancers often demonstrate indolent clinical behavior (*Miller et al.* (2006)), many men with low-risk tumors still receive surgery or radiation therapy, both of which are associated with potentially serious complications including incontinence, impotence, and other side effects (*Anandadas et al.* (2011)). These complications are particularly distressing given that evidence shows that these men may not survive longer with surgery or radiation than they do with expectant management approaches. Active Surveillance is a form of expectant management that involves monitoring patients by conducting regular clinical exams, biomarker tests, radiologic imaging, and biopsies. Due to concerns that many men who are diagnosed with prostate cancer are overtreated, active surveillance has been promoted as a way for low-risk men to delay and possibly avoid surgery or radiation treatment. However, many approaches to implementing active surveillance have been recommended and the best approach is unclear (*DallEra et al.* (2012)).

Due to a lack of evidence in support of a single optimal active surveillance strat-

egy, it is left to individual urologists and patients to decide how frequently to conduct follow-up biopsies. No previous study has made a link between different active surveillance follow-up strategies and the delay in the detection of progression to high-grade cancer. Risk of progression is one of the most important considerations when weighing long-term risk for patients on active surveillance. The ideal strategy to minimize risk of delaying the detection of high-grade cancer progression is to biopsy patients frequently (e.g. annually as suggested by *Tosoian et al.* (2011)). However, this risk competes with the harms of frequent biopsies resulting in pain and anxiety for patients, and the potential for complications such as infection. Infection rates for biopsy are approximately 1-2% (*Gonzalez et al.* (2012)); however, recent studies suggest that infection rates for patients undergoing active surveillance increases as a function of the number of biopsies they have received (*Ehdaie et al.* (2014)). Studies have also observed discontinuation of active surveillance by patients without signs of progression (*Loeb et al.* (2015)) and some have suggested that reducing surveillance biopsies may encourage compliance with active surveillance (*Al Otaibi et al.* (2008))

We used a *hidden Markov model* to evaluate longitudinal data from the Johns Hopkins Active Surveillance study to estimate initial biopsy sampling error, biopsy accuracy, and the rate of progression from low to intermediate or high-grade prostate cancer over time. Note that we use the term "progression" broadly to refer to biological progression of cancer grade and the occurrence of de novo cancer. We implemented a version of the Baum-Welch algorithm tailored to consider uncertainty in the initial population's health status due to biopsy sampling error. Next, we used sensitivity analysis based on simulated data to establish the algorithm converges to accurate parameter estimates for hidden Markov models. Finally, we used the model to evaluate all possible follow-up surveillance strategies as well as previously proposed strategies for active surveillance found in the literature on the basis of mean delay time to grade progression and the number of planned biopsies over the first 10 years following

initiation of active surveillance.

6.2 Model

In this section, we summarize the data from the Johns Hopkins Active Surveillance study; the hidden Markov model we used to estimate initial risk of the active surveillance cohort, biopsy sampling error, and prostate cancer grade progression; and how we used these estimates to develop a simulation model to evaluate alternative active surveillance strategies.

6.2.1 Data

The Johns Hopkins Active Surveillance study collected longitudinal data for men initially believed to have "favorable risk" prostate cancer. The data includes 1499 men with data collected over 20 years. The enrollment criteria were: clinical stage $\leq$ T1c, PSA density ≤ 0.15, Gleason score ≤ 6, total positive cores ≤ 2, and single core positivity $\leq 50\%$. Due to patient preference, older men with low-risk disease (i.e., clinical stage $\leq$ T2a, PSA < 10 ng/mL, and Gleason score ≤ 6) were also enrolled in the study. The data collected includes PSA, age, and biopsy results (e.g., Gleason score, number of positive cores, maximum percentage core involvement). The dataset used was anonymized with respect to patient identifiers and approval of the University of Michigan IRB was obtained prior to initiation of the study.

The original dataset included longitudinal data for 1521 patients. We removed 22 patients from the dataset due to missing diagnostic biopsy information. Table 6.1 describes the patient characteristics at diagnosis of the 1499 patients included in the study. Among men who discontinue active surveillance and receive treatment, 50.9% received surgery and 46.2% received radiation therapy. The study protocol called for patients to be biopsied annually. The mean and variance of the time between biopsies was 14.2 and 60.1 months, respectively. The median number of biopsies per

patient, including diagnosis biopsy, was 3 and ranged from 1 to 14. Table 6.2 shows the biopsy characteristics, where we have defined progression to be transition from a Gleason score ≤ 6 to Gleason score ≥ 7 on biopsy. Due to this definition, we excluded the six patients diagnosed with Gleason 7 disease from the analysis in Table 6.2.

Table 6.1: Patient characteristics at time of diagnosis. AS = active surveillance.

Characteristic	**AS cohort** ($N = 1499$), no. (%)
Age at diagnosis, yr	
≤ 49	18 (1.2)
50–59	208 (13.9)
60–69	911 (60.8)
70–79	352 (23.5)
≥ 80	10 (0.7)
Race	
White	1314 (87.7)
Black	115 (7.7)
Other	60 (4.0)
NA	10 (0.7)
PSA at diagnosis, ng/mL	
0–2.5	162 (10.8)
2.5–4	249 (16.6)
4–6	558 (37.2)
6–10	322 (21.5)
> 10	85 (5.7)
NA	123(8.2)
PSA density at diagnosis	
0–0.05	166 (11.1)
0.05–0.10	538 (35.9)
0.10–0.15	428 (28.6)
0.15–0.20	134 (8.9)
> 0.20	114 (7.6)
NA	119 (7.9)
Gleason score at diagnosis	
≤ 6	1488 (99.3)
$3 + 4$	5 (0.3)
$4 + 3$	1 (0.1)
NA	5 (0.3)

Table 6.2: Biopsy characteristics at diagnosis and surveillance biopsies.

		Biopsy						
Characteristics	**Diagnosis**	**First**	**Second**	**Third**	**Fourth**	**Fifth**	**Sixth**	**Seventh**
Patients, n	1493	1370	922	644	447	298	187	122
Age at biopsy, yr, mean (SD)	66 (6.0)	67 (6.1)	67 (6.0)	68 (5.5)	68 (5.3)	69 (5.1)	70 (5.1)	71 (4.3)
Months since last biopsy, mean (SD)	0 (0.0)	13 (8.2)	15 (7.5)	15 (7.1)	16 (8.6)	15 (6.9)	16 (7.6)	14 (3.9)
Most recent PSA, ng/mL, mean (SD)								
Most recent PSA density, mean (SD)	0.12 (0.07)	0.11 (0.07)	0.11 (0.08)	0.10 (0.07)	0.10 (0.09)	0.10 (0.07)	0.10 (0.08)	0.09 (0.07)
No. of biopsy cores, median (range)	12 (6–58)	12 (4–31)	12 (6–60)	12 (6–28)	12 (8–18)	12 (6–16)	14 (6–24)	14 (6-15)
Percentage of cores positive for cancer, n (%)								
0	0 (0.0)	568 (41.5)	435 (47.2)	336 (52.2)	237 (53.0)	154 (51.7)	97 (51.9)	57 (46.7)
> 0 and $< 34\%$	808 (54.1)	624 (45.5)	415 (45.0)	271 (42.1)	188 (42.1)	129 (43.3)	79 (42.2)	58 (47.5)
$\geq 34\%$	10 (0.7)	58 (4.2)	20 (2.2)	7 (1.1)	9 (2.0)	5 (1.7)	5 (2.7)	2 (1.6)
NA	675 (45.2)	120 (8.8)	52 (5.6)	30 (4.7)	13 (2.9)	10 (3.4)	6 (3.2)	5 (4.1)
Gleason score, n (%)								
No cancer	0 (0.0)	568 (41.5)	435 (47.2)	336 (52.2)	237 (53.0)	154 (51.7)	97 (51.9)	57 (46.7)
≤ 6	1488 (99.7)	670 (48.9)	413 (44.8)	275 (42.7)	181 (40.5)	130 (43.6)	78 (41.7)	54 (44.3)
7 (3+4)	0 (0.0)	78 (5.7)	49 (5.3)	16 (2.5)	18 (4.0)	9 (3.0)	10 (5.3)	5 (4.1)
7 (4+3)	0 (0.0)	30 (2.2)	14 (1.5)	12 (1.9)	6 (1.3)	4 (1.3)	1 (0.5)	3 (2.5)
≥ 8	0 (0.0)	18 (1.3)	7 (0.8)	1 (0.2)	2 (0.4)	1 (0.3)	0 (0.0)	1 (0.8)
NA	5 (0.3)	6 (0.4)	4 (0.4)	4 (0.6)	3 (0.7)	0 (0.0)	1 (0.5)	2 (1.6)
Outcome, n (%)								
Progression	0 (0.0)	126 (9.2)	70 (7.6)	29 (4.5)	26 (5.8)	14 (4.7)	11 (5.9)	9 (7.4)
No progression	1493 (100.0)	1244 (90.8)	852 (92.4)	615 (95.5)	421 (94.2)	284 (95.3)	176 (94.1)	113 (92.6)

6.2.2 Hidden Markov Model for Prostate Cancer Grade Progression

The specific type of model we employ is a hidden Markov model in which patient's progress through health states as defined by their prognostic grade groups based on Gleason score, the most important clinical factor for assessing risk of prostate cancer mortality. The term hidden refers to the fact that the exact health state of the patient is unknown in the absence of prostatectomy. The probability of progression to a higher prognostic grade group is determined by transition probabilities. We based the model on one-year time periods between state transitions to be consistent with the highest proposed frequency of biopsies and because that was the planned frequency of biopsies in the Johns Hopkins study.

In the remainder of this section, we use the notation from *Rabiner* (1989) to describe the hidden Markov model. We index annual time periods as $t = 0, 1, \ldots, T$. The model has states at time period t denoted by $s_t \in S \equiv \{S_0, S_1\}$, where S_0 denotes patients with low-grade cancer (defined as Gleason score ≤ 6), and S_1 denotes patients whose cancer has progressed to a higher grade cancer (defined as Gleason score ≥ 7). Since patients in the high-grade state cannot regress to the low-grade state the transition probability matrix is that of an absorbing Markov chain:

$$A = \begin{bmatrix} P(S_0|S_0) & P(S_1|S_0) \\ 0 & 1 \end{bmatrix}$$

These states are not directly observable. At $t = 0$, patients begin active surveillance under the belief that they are in state S_0; however, due to biopsy sampling error they could be in state S_1. We let $\pi = (\pi_0, \pi_1)$ denote the initial distribution of patients in states S_0 and S_1 at their first surveillance biopsy. Biopsies are performed at each time period (annually). The model has observations $o \in O \equiv \{O_0, O_1\}$ where O_0 denotes a biopsy observation of Gleason score ≤ 6 and O_1 denotes a biopsy observation of Gleason score ≥ 7 . Biopsies are imperfect due to sampling error and the

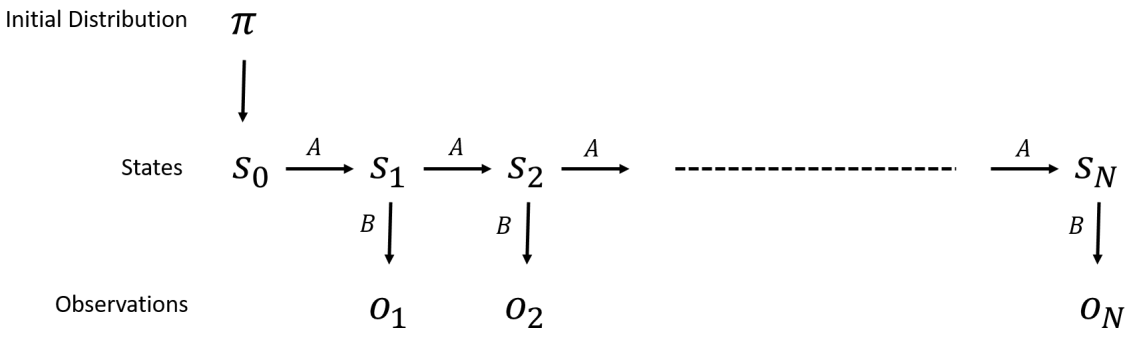

Figure 6.1: Illustration of the state transition and observation process for the hidden Markov model.

conditional probability of biopsy observations are denoted by the following matrix:

$$B = \begin{bmatrix} P(O_0|S_0) & P(O_1|S_0) \\ P(O_0|S_1) & P(O_1|S_1) \end{bmatrix}$$

If a biopsy result is O_1 (Gleason score ≥ 7), the patient exits the system and receives treatment. Collectively the model parameters for the hidden Markov model are denoted by $\lambda = (\pi, A, B)$. Figure 6.1 illustrates the stochastic active surveillance process.

We used the Baum-Welch algorithm to compute maximum likelihood estimates for the hidden Markov model parameters (*Rabiner* (1989)). The Baum-Welch algorithm is a special case of the general expectation-maximization (EM) algorithm (*Dempster et al.* (1977)), an iterative algorithm that combines forward and backward passes on a longitudinal observation sequence to find the choice of λ that maximizes the likelihood of observing the collection of sequences. In our application, we have biopsy results for a collection of $k = 1, \ldots, N$ patients. Each patient, k, results in an observation sequence, $O^{(k)} = [O_1^{(k)}, O_2^{(k)} \ldots O_{T_k}^{(k)}]$, which represents a patient's biopsy results over T_k time periods. We denote the set of N observation sequences as $O =$

$[O^{(1)}, O^{(2)}, \ldots, O^{(N)}]$. Thus, our goal is to find the λ that maximizes:

$$P(O|\lambda) = \prod_{k=1}^{N} P(O^{(k)}|\lambda) = \prod_{k=1}^{N} P_k \tag{6.1}$$

where we assume that observation sequences between patients are independent.

To describe the Baum-Welch algorithm, we define some additional parameters. We denote elements of matrices A and B as a_{ij} and b_{ij}, respectively. First, we define the forward variable $\alpha_t^k(i)$ as:

$$\alpha_t^k(i) = P(O_1^{(k)}, O_2^{(k)}, \ldots, O_t^{(k)}, s_t = S_i|\lambda)$$

which is the probability of observing the partial observation sequence, $O_1^{(k)}, O_2^{(k)}, \ldots O_t^{(k)}$, (until time t) and being in state S_i at time t, given the model λ. We use forward induction to solve for $\alpha_t^k(i)$:

$$\alpha_1^k(i) = \pi_i b_i(O_1^{(k)}), \quad i = 1, 2$$

$$\alpha_{t+1}^k(j) = \left[\sum_{i=1}^{N} \alpha_t^k(i) a_{ij}\right] b_j\big(O_{t+1}^{(k)}\big), \quad 2 \leq t \leq T-1, \ j = 1, 2$$

Next, we define the backward variable $\beta_t^k(i)$ as

$$\beta_t^k(i) = P(O_{t+1}^{(k)}, O_{t+2}^{(k)}, \ldots, O_{T_k}^{(k)}|\lambda, s_t = S_i)$$

which is the probability of the partial observation sequence from $t+1$ to T_k, given the model λ and given the patient is in state S_i at time t. We use backward induction to solve for $\beta_t^k(i)$:

$$\beta_T^k(i) = 1, \quad i = 1, 2$$

$$\beta_t^k(i) = \sum_{j=1}^{N} a_{ij} b_j\big(O_{t+1}^{(k)}\big)\beta_{t+1}^k(j), \quad t = T-1, T-2, \ldots, 1, \ i = 1, 2$$

To define the iterative procedure, *Rabiner* (1989) defines $\xi_t^k(i,j)$ to be the probability of patient k being in state S_i at time t and state S_j at time $t+1$ given the model λ and observation sequences of patient k, $O^{(k)}$:

$$\xi_t^k(i,j) = P(s_t = S_i, s_{t+1} = S_j | O^{(k)}, \lambda) = \frac{\alpha_t^k(i) a_{ij} b_j(O_{t+1}^{(k)}) \beta_{t+1}(j)}{P_k}.$$

$\gamma_t^k(i)$ is the probability of patient k being in state S_i at time t, given the model and the observation sequences of patient k, $O^{(k)}$. Thus, we can relate $\gamma_t^k(i)$ and $\xi_t^k(i,j)$:

$$\gamma_t^k(i) = \sum_{j=1}^{N} \xi_t^k(i,j)$$

Based on these definitions we can write the following update formulas, which iteratively improve $P(O|\lambda)$ using forward-backward equations:

$$\overline{a_{ij}} = \frac{\sum_{k=1}^{N} \frac{1}{P_k} \sum_{t=1}^{T_k - 1} \xi_t^k(i,j)}{\sum_{k=1}^{N} \frac{1}{P_k} \sum_{t=1}^{T_k - 1} \gamma_t^k(i)} \tag{6.2}$$

$$\overline{b_j}(l) = \frac{\sum_{k=1}^{N} \frac{1}{P_k} \sum_{t=1 \ s.t. \ O_t^{(k)} = l}^{T_k - 1} \gamma_t^k(j)}{\sum_{k=1}^{N} \frac{1}{P_k} \sum_{t=1}^{T_k - 1} \gamma_t^k(j)} \tag{6.3}$$

$$\overline{\pi_i} = \sum_{k=1}^{N} \frac{\gamma_1^k(i)}{P_k} \tag{6.4}$$

The update equation for $\overline{a_{ij}}$ calculates the expected number of transitions from state S_i to state S_j divided by the expected number of transitions from state S_i. The update equation for $\overline{b_j}(l)$ calculates the expected number of times a patient is in state S_j and observes l divided by the expected number of times a patient is in state S_j. Finally, the update equation for $\overline{\pi_i}$ is the expected number of times a patient is in

state S_i at $t = 1$. The Baum-Welch algorithm uses the above formulas to iteratively update the parameters of λ. A proof of convergence follows from the convergence guarantees for the EM algorithm (*Rabiner* (1989)). However, convergence to a local optimum is possible since the maximization problem is not strictly convex, and thus the limiting point for the sequential updates may be sensitive to the starting point. For this reason we conducted sensitivity analysis using data generated by sampling from known models to confirm convergence of the Baum-Welch algorithm.

Algorithm 2 Baum-Welch algorithm for parameter estimation.

Input: Initial model parameter estimates $\lambda^0 = (A^0, B^0, \pi^0)$.
1: Calculate $P(O|\lambda^0)$ using equation 6.1.
2: Calculate $\lambda^1 = (A^1, B^1, \pi^1)$, which is a function of λ^0 using the update equations 6.2, 6.3, and 6.4.
3: Calculate $P(O|\lambda^1)$ using equation 6.1.
4: $v \leftarrow 1$
5: **while** $P(O|\lambda^v) - P(O|\lambda^{v-1}) > 10^{-6}$ **do**
6: $\quad v \leftarrow v + 1$
7: $\quad$ Calculate $\lambda^v = (A^v, B^v, \pi^v)$, which is a function of λ^{v-1} using equations 6.2, 6.3, and 6.4.
8: $\quad$ Calculate $P(O|\lambda^v)$.
9: **end while**

6.2.3 Model Validation

To further validate the results obtained, we used the base case estimates of our model to simulate the detection rate based on 10,000 samples assuming annual biopsy as planned in the Johns Hopkins Active Surveillance study protocol, and compared the results to the observed detection rates in the Johns Hopkins data.

Next, we conducted experiments based on a hypothetical hidden Markov model for which we knew the true values for model parameters, and we tested our implementation of the Baum-Welch algorithm on sampled results for 1375 simulated patient observation sequences, which is the number of patients in the study who received their first surveillance biopsy. Since there was missing data in the Johns Hopkins Ac-

tive Surveillance study resulting from patients who discontinued active surveillance in the absence of grade progression, we sought to test the assumption that the missing data was not informative. Therefore, we censored the data for simulated observation sequences according to the observed mean rate of patients discontinuing active surveillance without grade progression. We then ran the Baum-Welch algorithm on the simulated data and compared the parameter estimates to the true parameters used to generate the simulated data.

6.2.4 Sensitivity Analysis

To validate that the resulting parameter estimates were not sensitive to the initial starting points, we varied our initial estimate for each parameter using a range of ± 0.1 with an upper limit of 0.99. We then ran the Baum-Welch algorithm on each new set of initial estimates, and compared the resulting parameter estimates.

Additionally, we performed bootstrapping analysis for which we randomly sampled 1375 patients with replacement from the Johns Hopkins dataset. We generated 30 different bootstrap samples and ran the Baum-Welch algorithm on each sample, and compared the resulting parameter estimates.

6.2.5 Simulation Model

The hidden Markov model can be used to compute statistical estimates of health outcomes such as the mean delay in detection of grade progression for patients who experienced grade progression. The delay time depends on the hidden Markov model parameter estimates, λ, which includes the initial probability a patient has Gleason Score ≤ 6 (i.e., π_0) or Gleason Score ≥ 7 (i.e., π_1) at the time of diagnosis, the transition probability from Gleason Score ≤ 6 to Gleason Score ≥ 7 (i.e., a_{01}), and the sensitivity (i.e., b_{11}) and specificity of biopsy (i.e., b_{00}). Together with the active surveillance biopsy schedule these parameters collectively govern the time to reach the

high-grade cancer state and subsequent detection of grade progression. We defined the biopsy schedule as a vector, $\vec{x} = (x_1, x_2, \ldots, x_T)$ of binary decision variables where $x_t = 1$ indicates a biopsy is planned at period t and $x_t = 0$ indicates that a biopsy is not planned at period t. The problem of determining the optimal active surveillance schedule can be expressed as follows:

$$\vec{x}^* = \underset{x}{\operatorname{argmin}} \left(\mathbb{E}[D(\omega, \vec{x})|\lambda] \;\middle|\; \sum_{i=1}^{T} x_i \leq \delta \right)$$

where $D(\omega, \vec{x})$ denotes the delay time for detection of progression for a given random sample path, ω, that indexes all the outcomes of the hidden Markov model including the true state at each time period, s_t, sampled using transition probability matrix A and the biopsy outcome, sampled using the observation matrix, B. The expectation is with respect to the hidden Markov model and can be estimated by random sampling. The parameter δ is a limit on the number of biopsies allowed over the T year time horizon. By varying δ, the set of Pareto optimal schedules can be obtained. The number of possible active surveillance strategies is 2^T and the optimization problem can be solved via total enumeration. We simulated all possible strategies and identified those strategies that were non-dominated, i.e., those strategies for which no other strategy simultaneously recommended fewer biopsies and had lower mean time to detect high-grade cancer.

6.3 Results

6.3.1 Hidden Markov Model Analysis

To initiate the Baum-Welch algorithm, we needed initial estimates of the model parameters $\lambda = (\pi, A, B)$. These estimates are not directly observable in the dataset, because biopsies are imperfect; thus, we used estimates from the literature. *Alam et al.* (2015) studied reclassification rates for men in the Johns Hopkins Active Surveillance

study, and found that the majority of men are reclassified within the first two years, most likely due to initial biopsy misclassification. We estimated annual progression rate from Gleason score ≤ 6 to Gleason score ≥ 7 to be 5% by calculating the rate of progression at patients' third through thirteenth biopsies. Estimates for the sensitivity and specificity of biopsy to Gleason score ≥ 7 disease were calculated to be 62.5% and 89.4%, respectively, based on data reported in *Epstein et al.* (2012), which compared biopsy results to Gleason score at radical prostatectomy. Finally, using data reported in *Epstein et al.* (2012), we estimated that 74.9% of patients diagnosed with Gleason score ≤ 6 disease on biopsy have Gleason score ≤ 6 disease at radical prostatectomy, while 25.1% have Gleason score ≥ 7 disease at radical prostatectomy. Based on these estimates from the literature, we used the following parameter starting points to initiate the Baum-Welch algorithm:

$$A = \begin{bmatrix} 0.950 & 0.050 \\ 0 & 1 \end{bmatrix}, \; B = \begin{bmatrix} 0.894 & 0.106 \\ 0.375 & 0.625 \end{bmatrix}, \; \pi = \begin{bmatrix} 0.749 & 0.252 \end{bmatrix}$$

After estimating these initial parameter estimates, we ran the Baum-Welch algorithm with a stopping criteria defined by a tolerance of 10^{-6} on the difference between the log likelihoods for consecutive iterations. The resulting parameter estimates from the Baum-Welch algorithm were:

$$\tilde{A} = \begin{bmatrix} 0.960 & 0.040 \\ 0 & 1 \end{bmatrix}, \; \tilde{B} = \begin{bmatrix} 0.986 & 0.014 \\ 0.390 & 0.610 \end{bmatrix}, \; \tilde{\pi} = \begin{bmatrix} 0.866 & 0.134 \end{bmatrix}$$

Thus, we found that annual progression rate from Gleason score ≤ 6 to Gleason score ≥ 7 to be 4%; the sensitivity and specificity of biopsy to Gleason score ≥ 7 disease to be 61.0% and 98.6%, respectively; and the initial proportion of patients undergoing active surveillance with Gleason score ≤ 6 disease at their first surveillance biopsy to be 86.6%.

Table 6.3: Results comparing hidden Markov model parameter estimates from the Baum-Welch algorithm to the true model parameter estimates from a known model.

Model Parameter	True Value	95% Confidence Interval
a_{00}	0.960	[0.961,0.967]
b_{00}	0.986	[0.980,0.988]
b_{11}	0.610	[0.582,0.688]
π_0	0.866	[0.846,0.889]

6.3.2 Validation

To validate the results obtained, we used the base case estimates of our model to simulate the detection rate of Gleason score ≥ 7 disease assuming annual biopsy as planned in the Johns Hopkins Active Surveillance study protocol. Figure 6.2 compares the model-based results and the observed results. Model predicted results were based on 10,000 samples. The confidence intervals for the observed results are reported in the figure. There was no statistically significant difference between the predicted and observed biopsy detection rate at the p=0.05 threshold.

Results for the hypothetical hidden Markov model for which the true values for model parameters are known are presented in Table 6.3, which shows the true model parameters from the known model and the 95% confidence interval for our model parameter estimates based on the Baum-Welch algorithm applied to the simulated sequences. The true value of π_0, b_{11}, and b_{00} all lie within the 95% confidence intervals, while the true value of a_{01} is 0.040 and does not lie within the 95% confidence interval, $0.033 - 0.039$.

6.3.3 Sensitivity Analysis

To validate that the resulting estimates are not sensitive to the starting points, for each parameter, we varied the initial estimates of the model parameters in the range of ± 0.1 with an upper limit of 0.99. We found the resulting parameter estimates

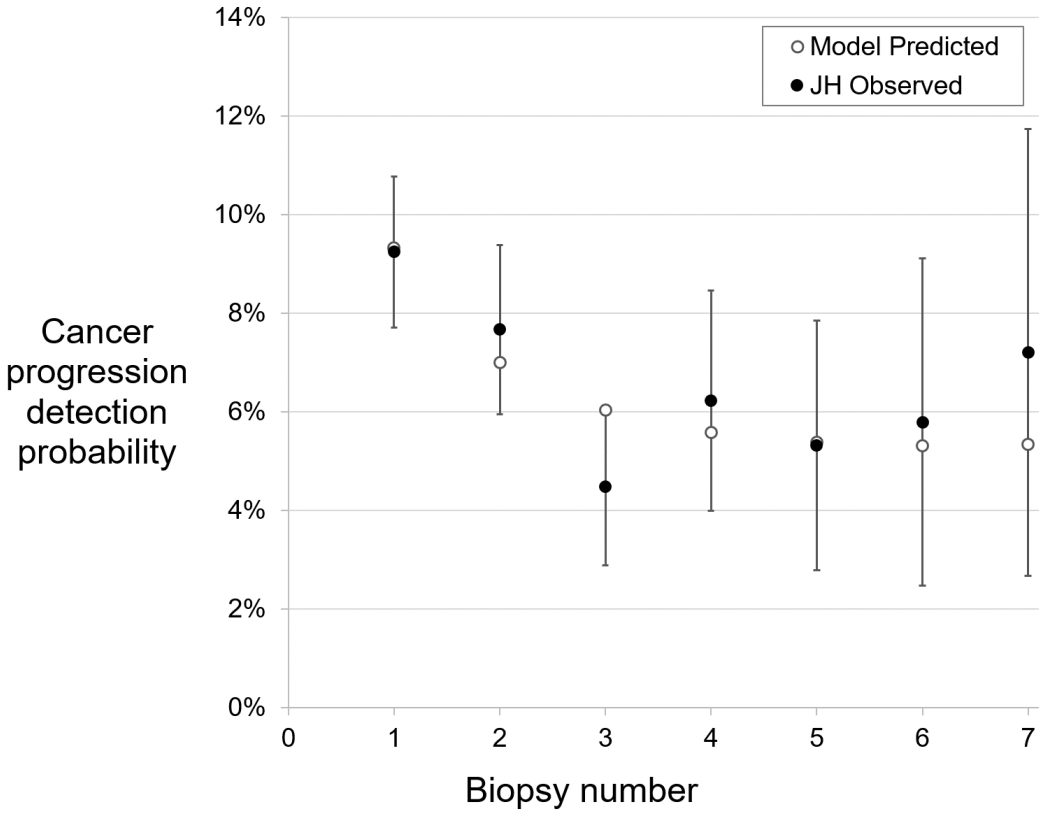

Figure 6.2: Comparison of Gleason score ≥ 7 detection rates predicted by the simulation model to the observed rate in the Johns Hopkins study. Model predicted results were based on 10,000 samples. The confidence intervals for the observed results are shown, and the confidence intervals for the model predicted results are too small to see on the figure.

Table 6.4: Bootstrapping results.

Parameter	Mean	95% Confidence Interval
a_{00}	0.964	[0.960, 0.968]
b_{00}	0.984	[0.980, 0.988]
b_{11}	0.611	[0.587, 0.635]
π_0	0.867	[0.857, 0.876]

varied by less than 0.5% from the values calculated using our original starting points, suggesting that starting points did not significantly impact our parameter estimates.

For our bootstrapping analysis we generated 30 different bootstrap samples and ran the Baum-Welch algorithm on each sample. The resulting 95% confidence intervals based on bootstrapping are presented in Table 6.4, with b_{11} having the most variation.

6.3.4 Optimization of Active Surveillance Strategies

The Baum-Welch algorithm calculated that 86.6% of patients were correctly diagnosed to have Gleason score ≤ 6 and 13.4% of patients were undersampled at their first surveillance biopsy and actually had Gleason score ≥ 7. In our simulation model we assumed diagnosis occurred 12 months prior to the first surveillance biopsy. Thus, based on hidden Markov model parameters from the Baum-Welch algorithm we estimated that 90.22% of patients had Gleason score ≤ 6 at diagnosis, and 9.78% of patients were undersampled and had Gleason score ≥ 7 at diagnosis.

We used the hidden Markov model parameter estimates to simulate mean delay time in detecting progression among patients who progress to high-grade cancer over a 10-year period following diagnosis of prostate cancer for all 2^{10} possible active surveillance strategies. Our simulation model found that 40% of patients would progress to high-grade cancer in 10 years, and that a strategy that performs annual biopsies (the Johns Hopkins strategy) takes a mean of 14.1 months to detect progression. The strategy that minimized the mean delay time for each choice of planned number of

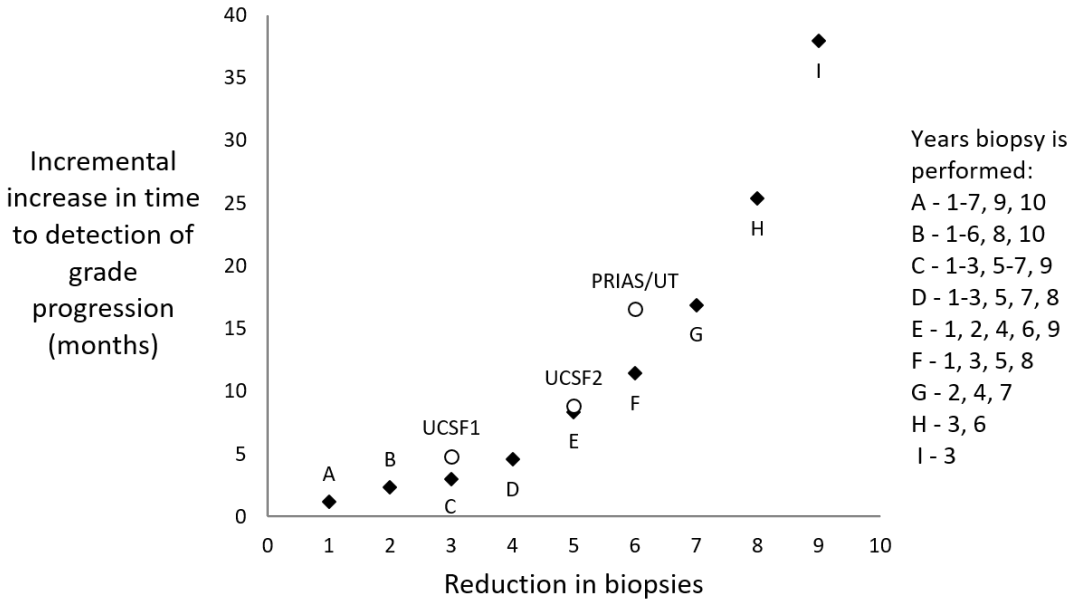

Figure 6.3: Simulation results for optimal active surveillance strategies and published strategies based on the estimated hidden Markov model parameters. Incremental time to detection and the reduction in biopsies are relative to an annual biopsy strategy. (Note: mean time to detection of grade progression for annual biopsy plan = 14.1 months)

biopsies over 10 years are plotted in Figure 6.3, which shows the incremental time to detection and the reduction in biopsies relative to a strategy that performs annual biopsies. We have also included active surveillance strategies from the literature. University of California, San Francisco (UCSF) recommends a biopsy 1 year after diagnosis, then every 1 to 2 years. We modeled two versions of this policy: UCSF1 performs a biopsy after 1 year, then every 1.5 years and UCSF2 performs a biopsy after 1 year, then every 2 years. PRIAS/University of Toronto (UT) performs a biopsy after 1, 4, 7, and 10 years (*van den Bergh et al.* (2007)). Figure 6.3 displays that UCSF performs well compared to the optimal policies, while the PRIAS/UT policy increases mean time to detection by 5.2 months compared to our optimal policy that performs biopsies after 1, 3, 5, and 8 years.

6.4 Conclusions

Many experts have called for the use of active surveillance to address overtreatment concerns for men with low-risk prostate cancer. active surveillance delays and possibly avoids immediate treatment via surgery or radiation therapy until and unless there is evidence that the disease has progressed; however, it comes with a burden to patients due to the need to conduct follow-up clinical exams, tests, and surveillance biopsies. Biopsies in particular are a significant burden to patients. The intensiveness of follow-up determines the frequency of clinical exams, tests, and biopsies. In the absence of randomized trials comparing active surveillance pathways there is no consensus among urologists about the best way to trade-off the burden of surveillance with the benefits of avoiding cancer progression (*Lawrentschuk and Klotz* (2011)). Furthermore, little is known about the factors that determine why men initially choose active surveillance over immediate treatment (surgery or radiation therapy), or why they choose to abandon active surveillance. We provide a new active surveillance precision medicine (ASPM) model for quantifying the trade-off between benefits and harms of various active surveillance strategies. These decisions must trade-off between the potential long-term benefits of detecting disease progression with the burden of surveillance, including patient anxiety, and the potential harms and side effects of biopsies (e.g. pain, anxiety, and hospitalization for infection in 2–3% of cases).

Our results suggest that there are diminishing benefits as the number of biopsies increases. It is interesting to note that we can eliminate six biopsies in 10 years, which has a substantial impact on the patient, while only increasing the time to detection by 11.4 months. Additionally, the optimal biopsy schedules tend to perform more biopsies in the beginning to catch patients misdiagnosed with low-risk cancer due to undersampling of biopsies.

There are multiple definitions of progression for prostate cancer, including definitions based on increase in PSA, PSA velocity and density, and tumor volume. Grade

progression, which refers to a change in Gleason score, is a definitive form of progression recognized by all published active surveillance guidelines. However, currently there is debate about whether grade progression is possible, or if the occurrence of higher grade cancer on biopsy occurs due to biopsy sampling error. Some studies suggest that that a combination of sampling error and progression are responsible for increased grade detection over time (*Epstein et al.* (2012); *Inoue et al.* (2014)). Our findings lend additional evidence to these studies, suggesting a combination of sampling error and progression are responsible for detection of higher grade cancers in the future.

A chief concern about active surveillance is the possibility that prostate cancer progresses in the interval of time between biopsies or that progression is missed due to imperfect sensitivity of biopsies. The potential for undetected progression raises questions about health outcomes for patients on active surveillance who progress and receive treatment. Studies comparing radical prostatectomy outcomes for patients initially on active surveillance to patients receiving radical prostatectomy immediately following diagnosis have shown that low-risk men who receive annual biopsy on active surveillance do not have worse surgery outcomes (*Tosoian et al.* (2011)). Additionally, *Klotz et al.* (2014) reported that patients undergoing active surveillance with biopsies every 3 to 4 years had mortality rates consistent with patients who received initial definitive treatment. Assuming a uniform distribution of progression times during the 3 to 4 years intervals would suggest delays of 1.5 to 2 years in detecting grade progression may not have a clinically significant impact.

The most related work to ours is an article by *Inoue et al.* (2014) in which the authors develop a statistical model based on the assumption that patients may progress prior, during, or following the active surveillance study period. The authors' approach assumes that the risk of progression is stationary and the change point (at which progression occurs) can be modeled as a multinomial distribution. Their model

is premised on the assumption that the misclassification error is known, or can be estimated, and the authors' results show that these estimates are critically important. Unfortunately such estimates are never available in practice and must be estimated from other study populations. Our proposed approach relaxes this assumption, providing estimates of misclassification error that can be used to inform the optimal strategy for active surveillance.

One limitation is that our results apply to patients with "favorable risk" prostate cancer (i.e., clinical stage $\leq$ T1c, PSA density ≤ 0.15, Gleason score ≤ 6, total positive cores ≤ 2, and single core positivity $\leq 50\%$) and older men with low-risk disease (i.e., clinical stage $\leq$ T2a, PSA < 10, and Gleason score ≤ 6), since these are the patients that were enrolled in the Johns Hopkins Active Surveillance study and thus there is a need to validate our findings on other active surveillance studies; however, this initial study lays the groundwork for such future validation work. A related limitation is that our study is based on a single active surveillance study. Missing data from patients that drop out of the study without progression could confound the results, although our sensitivity analysis helped to mitigate this concern. Finally, our results provide the trade-off between number of biopsies and mean delay time to detection of progression; however, the amount of time that is considered safe to delay detection is not known. Nevertheless, data from the literature suggests that short delay times may not have significant clinical impact. Metastases is a better endpoint, but the data needed to fit a hidden Markov model with this endpoint does not yet exist. These limitations notwithstanding, we believe this study provides important evidence about the trade-off between varying active surveillance strategies and the optimal timing of biopsies during active surveillance. In the future, we would like to use datasets from other active surveillance studies to study optimal biopsy strategies for intermediate-risk patients.

While annual biopsy for low-risk men on active surveillance is associated with the

shortest time to detection of Gleason ≥ 7 disease, several alternative strategies may allow for less frequent biopsy without sizable increases in time to detecting grade progression. It is interesting to note that by performing biopsies in 1, 3, 5, and 8 years after diagnosis, we can eliminate six biopsies in 10 years, which has a substantial impact on the patient's quality of life and risk of infection, while only increasing the time to detection by 11.4 months. Additionally, the optimal biopsy schedules tend to perform more biopsies in the beginning to catch patients misdiagnosed with low-risk cancer. Moreover, we found that while the UCSF policy performed almost as well as optimal policies with the same number of biopsies over a 10 year period, we could reduce the time to detection by an additional 5.2 months compared to the PRIAS/UT policy while performing the same number of biopsies in 10 years.

CHAPTER VII

Conclusions

Prostate cancer screening can improve patient outcomes by catching cancer at an early stage when health outcomes are most favorable for patients; however, widespread prostate cancer screening has led to unnecessary biopsies caused by false-positive PSA results and the overtreatment of low-risk prostate cancer with harsh curative treatments. In this dissertation we presented a series of prostate cancer models to evaluate the use of new technologies for prostate cancer screening to better select men for prostate biopsy, as well as the optimal timing of surveillance biopsies for men with low-risk disease undergoing active surveillance. Following is a summary of the most important findings from Chapters III, IV, V, and VI.

In Chapter III, we developed and validated a new partially observable Markov model that considers prostate cancer screening and treatment decisions for a cohort of men, starting at age 40, through to end of life. We used this model to examine alternative choices of two-stage biomarker-based screening strategies based on newly discovered biomarkers with varying thresholds. The screening strategy with a PSA threshold of 2 ng/mL and a second biomarker with high-grade sensitivity and specificity of 0.86 and 0.62, respectively, increased the number of QALYs per 1000 men by 19 QALYs compared to no screening and by 7 QALYs compared to using the PSA test alone with a threshold of 4 ng/mL. Our model predicts one prostate

cancer death averted per 200 men screened. In our one-way sensitivity analysis, we found that other-cause mortality had the greatest impact on the expected increase in QALYs relative to no screening, suggesting that the presence of comorbidity is an important consideration when determining the optimal prostate cancer screening strategy.

In our analysis, we found that many different screening strategies performed similarly in terms of QALYs; however, it is possible to distinguish these similar screening strategies by looking at additional performance measures that may better account for patient preferences. For example, some strategies that achieved similar QALYs varied significantly in rates of biopsy and prostate cancer deaths, with reductions in prostate cancer deaths coming at the expense of a greater biopsy rate. This trade-off emphasizes the importance of a shared decision making approach to account for patient preferences regarding risk of prostate cancer mortality and harms from biopsy.

Identifying biomarkers and risk thresholds optimized for identification of high-grade cancers had the greatest impact on measures of performance in the screening setting. Combining new biomarkers with PSA has the potential to reduce the number of screening biopsies (thus decreasing overdiagnosis) and decrease the rate of prostate cancer mortality. The sensitivity analysis suggests our conclusions are robust with respect to plausible variation in model parameters.

In Chapter IV, we developed a Markov model to evaluate the cost-effectiveness of using MRI in a screening setting. We estimated the number of prostate cancer deaths averted, QALYs, and total cost for each strategy. Additionally, we estimated the ICERs. Interestingly, the strategies that performed a standard biopsy on negative MRI were more expensive and less effective than strategies that perform no biopsy on negative MRI. Based on our study, MRI as an intermediate test in the screening of men for prostate cancer is cost-effective assuming a willingness-to-pay threshold of \$100,000/QALY threshold. The most efficient strategy was the use of MRI if

PSA $>$ 4 ng/mL, followed by combined biopsy if MRI was positive and no biopsy if MRI was negative, using a PI-RADS threshold of 3 to indicate a positive MRI (ICER: \$23,483/QALY gained). These results were robust over a range of sensitivity analyses and were maintained even if the sensitivity and specificity of MRI and combined biopsy were reduced by 19 percentage points. This helps to establish the viability of MRI in a non-academic medical center setting where radiologists may be less experienced with MRI. Overall, our findings suggest that MRI appears to be a viable approach for early detection of prostate cancer from a cost-effectiveness perspective.

In Chapter V, we developed a new POMDP model to investigate optimal prostate cancer screening decisions using new biomarkers based on a patient's *belief state* rather than making decisions based solely on a patient's most recent test results. The belief state is calculated using Bayesian updating and comprises a patient's complete history of biomarker test results. We solved an approximation of the POMDP, which maximized total expected QALYs, based on a data-driven sampling approach to create a set of grid points in the belief space, and evaluated the resulting policy using our simulation model. Although exact solutions to the POMDP were not possible, we showed the grid based approximation could be solved to optimality. We also found that it is possible to develop approximate POMDP solutions based on data about frequently visited parts of the belief space, and that basing screening decisions on belief estimates that use the complete history of biomarker results may improve screening.

In Chapter VI, we used a hidden Markov model to estimate initial biopsy sampling error, biopsy accuracy, and the rate of progression from low to intermediate or high-grade prostate cancer over time based on longitudinal data from the Johns Hopkins Active Surveillance study. We used maximum likelihood estimation based on the Baum-Welch algorithm to estimate the hidden Markov model parameters and sensitivity analysis to establish robustness of the results. We used the resulting model to evaluate all possible follow-up surveillance strategies as well as previously proposed

strategies for active surveillance found in the literature on the basis of mean delay time to grade progression and the number of planned biopsies over the first 10 years following initiation of active surveillance.

While annual biopsy for low-risk men on active surveillance is associated with the shortest time to detection of Gleason ≥ 7 disease, several alternative strategies may allow for less frequent biopsy without sizable increases in time to detecting grade progression. In particular, performing biopsies in 1, 3, 5, and 8 years after diagnosis reduces the number of planned biopsies in 10 years by 6 compared to an annual biopsy schedule, while only increasing the mean time to detection by 11.4 months. We found that while the UCSF policy performed almost as well as our optimal policies, we could reduce the time to detection by an additional 5.2 months compared to the PRIAS/UT policy while performing the same number of biopsies in 10 years.

Chapters III, IV, and V of this dissertation have limitations based on assumptions used in the modeling process. First, estimates of sensitivity and specificity for biomarkers can be dataset-dependent, as the estimates come from different datasets and, therefore, may have different biases; however, our analysis still provides useful insights into how the sensitivity and specificity of biomarkers impact long-term health outcomes. Second, there is the potential for bias in the data we used to estimate MRI results because the population used includes patients with previous negative biopsies in addition to biopsy-naïve patients; however, by using the estimates based on the larger patient population we were able to obtain better estimates of sensitivity and specificity. Our sensitivity analysis further confirms our conclusions are not sensitive to this assumption.

Another potential limitation is that we assumed each patient receives at most one screening biopsy in his life. About $7 - 12\%$ of men undergoing biopsy have had a previous negative biopsy (*Nguyen et al.* (2010); *Thompson et al.* (2006)); however, the majority of patients receive a single biopsy, and cancers detected on second biopsy are

typically less clinically significant. Since our intent is to measure the public health impact of prostate cancer screening, we do not believe this assumption significantly influenced our results. Another possible limitation is the inconsistent definition of clinically significant prostate cancer in the literature. For example, *Siddiqui et al.* (2015) defined clinically significant disease as high-volume Gleason 3+4 or Gleason $\geq 4+3$, while *Grey et al.* (2015) defined clinically significant disease to be cancer core involvement ≥ 6 mm or the presence of any Gleason pattern 4. In our model, we considered clinically significant disease to be Gleason score ≥ 7. Finally, the only curative treatment included in our model was radical prostatectomy, because it is the most common curative treatment, and patients undergoing radiation therapy have similar health outcomes (*Hamdy et al.* (2016)).

One potential limitation of Chapter VI is that our results apply to patients with "favorable risk" prostate cancer (i.e., clinical stage $\leq$ T1c, PSA density ≤ 0.15, Gleason score ≤ 6, total positive cores ≤ 2, and single core positivity $\leq 50\%$) and older men with low-risk disease (i.e., clinical stage $\leq$ T2a, PSA < 10, and Gleason score ≤ 6), since these are the patients that were enrolled in the Johns Hopkins Active Surveillance study and thus there is a need to validate our findings on other active surveillance studies; however, this initial study lays the groundwork for such future validation work using other populations with different risk profiles. A related limitation is that our study is based on a single active surveillance study. Missing data from patients that drop out of the study without progression could confound the results, although our sensitivity analysis helps to mitigate this concern. Finally, our results provide the trade-off between number of biopsies and mean delay time to detection of progression; however, the amount of time that is considered safe to delay detection is not known. Nevertheless, data from the literature suggests that short delay times may not have significant clinical impact.

There are several possible research extensions from this dissertation. In our work

we have shown that new biomarkers provide clinical benefits to patients when used in conjunction with the PSA test; however, in the future it would be valuable to evaluate the cost-effectiveness of these new biomarkers for prostate cancer screening. Additionally, we evaluated the cost-effectiveness of MRI for prostate cancer screening for men from ages 55 to 69 with two year screening intervals; however, it would be interesting to further investigate the impact of screening ages on the cost-effectiveness of MRI. Furthermore, we studied the optimal use of new biomarkers and MRI in patients with elevated PSA with a fixed PSA threshold. Since PSA naturally increases as a patient ages, it would be beneficial to evaluate age-dependent PSA thresholds to trigger these secondary tests. Finally, it would be beneficial to validate our active surveillance precision medicine (ASPM) model on other cohorts of patients including intermediate risk patients, which would allow our results to be informative for a larger portion of the patient population.

In conclusion, we have provided important insights into how new technologies, like molecular biomarkers and MRI, can be used to supplement the PSA test for the early detection of prostate cancer, as well as the optimal timing of prostate biopsies for men with low-risk prostate cancer undergoing active surveillance. By using new technologies to better select men for biopsy and by improving active surveillance strategies, physicians can reduce the harms of prostate cancer screening (e.g., unnecessary biopsies and overtreatment of low-risk disease) while continuing to reduce prostate cancer deaths through screening and early detection. Finally, the work presented in this thesis could help lay the groundwork for early detection and treatment of other cancers.

BIBLIOGRAPHY

BIBLIOGRAPHY

Adibi, M., M. S. Pearle, and Y. Lotan (2012), Cost-effectiveness of standard vs intensive antibiotic regimens for transrectal ultrasonography (TRUS)-guided prostate biopsy prophylaxis, *BJU International*, *110*(2b), E86–E91.

Ahmed, H. U., et al. (2017), Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study, *The Lancet*.

Aizer, A. A., et al. (2015), Cost implications and complications of overtreatment of low-risk prostate cancer in the United States, *Journal of the National Comprehensive Cancer Network*, *13*(1), 61–68.

Al Otaibi, M., et al. (2008), Role of repeated biopsy of the prostate in predicting disease progression in patients with prostate cancer on active surveillance, *Cancer*, *113*(2), 286–292.

Alam, R., H. B. Carter, P. Landis, J. I. Epstein, and M. Mamawala (2015), Conditional probability of reclassification in an active surveillance program for prostate cancer, *The Journal of Urology*, *193*(6), 1950–1955.

Anandadas, C. N., et al. (2011), Early prostate cancer–which treatment do men prefer and why?, *BJU International*, *107*(11), 1762–1768.

Andriole, G. L., et al. (2009), Mortality results from a randomized prostate-cancer screening trial, *New England Journal of Medicine*, *360*(13), 1310–1319.

Andriole, G. L., et al. (2012), Prostate cancer screening in the randomized prostate, lung, colorectal, and ovarian cancer screening trial: mortality results after 13 years of follow-up, *Journal of the National Cancer Institute*, *104*(2), 125–132.

Arias, E. (2010), United States life tables 2006, *National Vital Statistics Reports*, *58*(21), 1–40.

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

Ayer, T., O. Alagoz, N. K. Stout, and E. S. Burnside (2015), Heterogeneity in women's adherence and its role in optimal breast cancer screening policies, *Management Science*, *62*(5), 1339–1362.

Barentsz, J. O., J. Richenberg, R. Clements, P. Choyke, S. Verma, G. Villeirs, O. Rouviere, V. Logager, and J. J. Fütterer (2012), ESUR prostate MR guidelines 2012, *European Radiology*, *22*(4), 746–757.

Birnbaum, J. K., Z. Feng, R. Gulati, J. Fan, Y. Lotan, J. T. Wei, and R. Etzioni (2015), Projecting benefits and harms of novel cancer screening biomarkers: a study of PCA3 and prostate cancer, *Cancer Epidemiology, Biomarkers & Prevention*, *24*(4), 677–682.

Bratt, O., and H. Lilja (2015), Serum markers in prostate cancer detection, *Current Opinion in Urology*, *25*(1), 59.

Brenner, J. C., A. M. Chinnaiyan, and S. A. Tomlins (2013), ETS fusion genes in prostate cancer, in *Prostate Cancer*, pp. 139–183, Springer.

Bryant, R. J., and H. Lilja (2014), Emerging PSA-based tests to improve screening, *Urologic Clinics of North America*, *41*(2), 267–276.

Bussemakers, M. J., A. van Bokhoven, G. W. Verhaegh, F. P. Smit, H. F. Karthaus, J. A. Schalken, F. M. Debruyne, N. Ru, and W. B. Isaacs (1999), DD3:: A new prostate-specific gene, highly overexpressed in prostate cancer, *Cancer Research*, *59*(23), 5975–5979.

Carter, H. B., et al. (2013), Early detection of prostate cancer: AUA guideline, *The Journal of Urology*, *190*(2), 419–426.

Cassandra, A. R., L. P. Kaelbling, and M. L. Littman (1994), Acting optimally in partially observable stochastic domains, in *AAAI*, vol. 94, pp. 1023–1028.

Catalona, W. J., et al. (2011), A multicenter study of [-2] pro-prostate specific antigen combined with prostate specific antigen and free prostate specific antigen for prostate cancer detection in the 2.0 to 10.0 ng/ml prostate specific antigen range, *The Journal of Urology*, *185*(5), 1650–1655.

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

Cooperberg, M. R., et al. (2011), Outcomes of active surveillance for men with intermediate-risk prostate cancer, *Journal of Clinical Oncology*, *29*(2), 228–234.

DallEra, M. A., et al. (2012), Active surveillance for prostate cancer: a systematic review of the literature, *European Urology*, *62*(6), 976–983.

de Koning, H. J., et al. (2014), Benefits and harms of computed tomography lung cancer screening strategies: a comparative modeling study for the US Preventive Services Task Force, *Annals of Internal Medicine*, *160*(5), 311–320.

de Rooij, M., S. Crienen, J. A. Witjes, J. O. Barentsz, M. M. Rovers, and J. P. Grutters (2014), Cost-effectiveness of magnetic resonance (MR) imaging and mr-guided targeted biopsy versus systematic transrectal ultrasound–guided biopsy in diagnosing prostate cancer: a modelling study from a health care perspective, *European Urology*, *66*(3), 430–436.

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

Draisma, G., R. Boer, S. J. Otto, I. W. van der Cruijsen, R. A. Damhuis, F. H. Schröder, and H. J. de Koning (2003), Lead times and overdetection due to prostate-specific antigen screening: estimates from the European randomized study of screening for prostate cancer, *Journal of the National Cancer Institute*, *95*(12), 868–878.

Draisma, G., R. Etzioni, A. Tsodikov, A. Mariotto, E. Wever, R. Gulati, E. Feuer, and H. de Koning (2009), Lead time and overdiagnosis in prostate-specific antigen screening: importance of methods and context, *Journal of the National Cancer Institute*, *101*(6), 374–383.

Ehdaie, B., et al. (2014), The impact of repeat biopsies on infectious complications in men with prostate cancer on active surveillance, *The Journal of Urology*, *191*(3), 660–664.

Epstein, J. I., Z. Feng, B. J. Trock, and P. M. Pierorazio (2012), Upgrading and downgrading of prostate cancer from biopsy to radical prostatectomy: incidence and predictive factors using the modified gleason grading system and factoring in tertiary grades, *European Urology*, *61*(5), 1019–1024.

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

Etzioni, R., D. F. Penson, J. M. Legler, D. di Tommaso, R. Boer, P. H. Gann, and E. J. Feuer (2002), Overdiagnosis due to prostate-specific antigen screening: lessons from US prostate cancer incidence trends, *Journal of the National Cancer Institute*, *94*(13), 981–990.

Etzioni, R., R. Gulati, S. Falcon, and D. F. Penson (2008a), Impact of psa screening on the incidence of advanced stage prostate cancer in the United States: a surveillance modeling approach, *Medical Decision Making*, *28*(3), 323–331.

Etzioni, R., et al. (2008b), Quantifying the role of PSA screening in the US prostate cancer mortality decline, *Cancer Causes & Control*, *19*(2), 175–181.

Ferro, M., et al. (2012), Predicting prostate biopsy outcome: prostate health index (phi) and prostate cancer antigen 3 (PCA3) are useful biomarkers, *Clinica Chimica Acta*, *413*(15), 1274–1278.

Gonzalez, C., T. Averch, L. Boyd, et al. (2012), AUA/SUNA white paper on the incidence, prevention and treatment of complications related to prostate needle biopsy, in *American Urological Association*.

Grann, V. R., P. R. Patel, J. S. Jacobson, E. Warner, D. F. Heitjan, M. Ashby-Thompson, D. L. Hershman, and A. I. Neugut (2011), Comparative effectiveness of screening and prevention strategies among BRCA1/2-affected mutation carriers, *Breast Cancer Research and Treatment*, *125*(3), 837–847.

Grey, A. D., M. S. Chana, R. Popert, K. Wolfe, S. H. Liyanage, and P. L. Acher (2015), Diagnostic accuracy of magnetic resonance imaging (MRI) prostate imaging reporting and data system (PI-RADS) scoring in a transperineal prostate biopsy setting, *BJU International*, *115*(5), 728–735.

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

Gulati, R., A. B. Mariotto, S. Chen, J. L. Gore, and R. Etzioni (2011), Long-term projections of the harm-benefit trade-off in prostate cancer screening are more favorable than previous short-term estimates, *Journal of Clinical Epidemiology*, *64*(12), 1412–1417.

Gulati, R., J. L. Gore, and R. Etzioni (2013), Comparative effectiveness of alternative prostate-specific antigen–based prostate cancer screening strategies model estimates of potential benefits and harms, *Annals of Internal Medicine*, *158*(3), 145–153.

Gulati, R., L. Y. Inoue, J. L. Gore, J. Katcher, and R. Etzioni (2014), Individualized estimates of overdiagnosis in screen-detected prostate cancer, *Journal of the National Cancer Institute*, *106*(2), djt367.

Haas, G. P., et al. (2007), Needle biopsies on autopsy prostates: Sensitivity of cancer detection based on true prevalence, *Journal of the National Cancer Institute*, *99*(19), 1484–1489.

Hamdy, F. C., et al. (2016), 10-year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer, *New England Journal of Medicine*, *375*(15), 1415–1424.

Hansen, N., et al. (2016), Magnetic resonance and ultrasound image fusion supported transperineal prostate biopsy using the ginsburg protocol: Technique, learning points, and biopsy results, *European Urology*, *70*(2), 332–340.

Hartigan, J. A., and M. A. Wong (1979), Algorithm as 136: A k-means clustering algorithm, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108.

Heijnsdijk, E., et al. (2015), Cost-effectiveness of prostate cancer screening: a simulation study based on ERSPC data, *Journal of the National Cancer Institute*, *107*(1), dju366.

Heijnsdijk, E. A., D. Denham, and H. J. de Koning (2016), The cost-effectiveness of prostate cancer detection with the use of prostate health index, *Value in Health*, *19*(2), 153–157.

Heijnsdijk, E. A., et al. (2012), Quality-of-life effects of prostate-specific antigen screening, *New England Journal of Medicine*, *367*(7), 595–605.

Howlader, N., et al. (2012), SEER cancer statistics review, 1975-2009 (vintage 2009 populations), national cancer institute, *Bethesda, MD*.

Inoue, L. Y., B. J. Trock, A. W. Partin, H. B. Carter, and R. Etzioni (2014), Modeling grade progression in an active surveillance study, *Statistics in Medicine*, *33*(6), 930–939.

Johansson, J.-E., O. Andrén, S.-O. Andersson, P. W. Dickman, L. Holmberg, A. Magnuson, and H.-O. Adami (2004), Natural history of early, localized prostate cancer, *Journal of the American Medical Association*, *291*(22), 2713–2719.

Klotz, L., D. Vesprini, P. Sethukavalan, V. Jethava, L. Zhang, S. Jain, T. Yamamoto, A. Mamedov, and A. Loblaw (2014), Long-term follow-up of a large active surveillance cohort of patients with prostate cancer, *Journal of Clinical Oncology*, *33*(3), 272–277.

Lavieri, M. S., M. L. Puterman, S. Tyldesley, and W. J. Morris (2012), When to treat prostate cancer patients based on their PSA dynamics, *IIE Transactions on Healthcare Systems Engineering*, *2*(1), 62–77.

Lawrentschuk, N., and L. Klotz (2011), Active surveillance for low-risk prostate cancer: an update, *Nature Reviews Urology*, *8*(6), 312–320.

Li, C.-k., B. C. Tong, and J. H. You (2016), Cost-effectiveness of culture-guided antimicrobial prophylaxis for the prevention of infections after prostate biopsy, *International Journal of Infectious Diseases*, *43*, 7–12.

Liu, J., P. R. Womble, S. Merdan, D. C. Miller, J. E. Montie, B. T. Denton, and M. U. S. I. Collaborative (2015), Factors influencing selection of active surveillance for localized prostate cancer, *Urology*, *86*(5), 901–905.

Loeb, S., H. B. Carter, S. I. Berndt, W. Ricker, and E. M. Schaeffer (2011), Complications after prostate biopsy: data from SEER-Medicare, *The Journal of Urology*, *186*(5), 1830–1834.

Loeb, S., Y. Folkvaljon, D. V. Makarov, O. Bratt, A. Bill-Axelson, and P. Stattin (2015), Five-year nationwide follow-up study of active surveillance for prostate cancer, *European Urology*, *67*(2), 233–238.

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

Makarov, D., S. Loeb, R. H. Getzenberg, and A. W. Partin (2009), Biomarkers for prostate cancer, *Annual Review of Medicine*, *60*, 139–51.

Mandelblatt, J. S., et al. (2016), Collaborative modeling of the benefits and harms associated with different US breast cancer screening strategies, *Annals of Internal Medicine*, *164*(4), 215–225.

Mariotto, A. B., K. Robin Yabroff, Y. Shao, E. J. Feuer, and M. L. Brown (2011), Projections of the cost of cancer care in the united states: 20102020, *Journal of the National Cancer Institute*.

Meng, X., et al. (2016), Relationship between prebiopsy multiparametric magnetic resonance imaging (MRI), biopsy indication, and MRI-ultrasound fusion–targeted prostate biopsy outcomes, *European Urology*, *69*(3), 512–517.

Merdan, S., P. R. Womble, D. C. Miller, C. Barnett, Z. Ye, S. M. Linsell, J. E. Montie, and B. T. Denton (2014), Toward better use of bone scans among men with early-stage prostate cancer, *Urology*, *84*(4), 793–798.

Merdan, S., S. A. Tomlins, C. L. Barnett, T. M. Morgan, J. E. Montie, J. T. Wei, and B. T. Denton (2015), Assessment of long-term outcomes associated with urinary prostate cancer antigen 3 and TMPRSS2:ERG gene fusion at repeat biopsy, *Cancer*, *121*(22), 4071–4079.

Miller, D. C., S. B. Gruber, B. K. Hollenbeck, J. E. Montie, and J. T. Wei (2006), Incidence of initial local therapy among men with lower-risk prostate cancer in the United States, *Journal of the National Cancer Institute*, *98*(16), 1134–1141.

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

Moyer, V. A. (2012), Screening for prostate cancer: US Preventive Services Task Force recommendation statement, *Annals of Internal Medicine*, *157*(2), 120–134.

Neumann, P. J., J. T. Cohen, and M. C. Weinstein (2014), Updating cost-effectiveness – the curious resilience of the $50,000-per-QALY threshold, *New England Journal of Medicine*, *371*(9), 796–797.

Nguyen, C. T., C. Yu, A. Moussa, M. W. Kattan, and J. S. Jones (2010), Performance of prostate cancer prevention trial risk calculator in a contemporary cohort screened for prostate cancer and diagnosed by extended prostate biopsy, *The Journal of Urology*, *183*(2), 529–533.

Oberlin, D. T., D. D. Casalino, F. H. Miller, R. S. Matulewicz, K. T. Perry, R. B. Nadler, S. Kundu, W. J. Catalona, and J. J. Meeks (2016), Diagnostic value of guided biopsies: fusion and cognitive-registration magnetic resonance imaging versus conventional ultrasound biopsy of the prostate, *Urology*, *92*, 75–79.

Parekh, D. J., et al. (2015), A multi-institutional prospective trial in the USA confirms that the 4Kscore accurately identifies men with high-grade prostate cancer, *European Urology*, *68*(3), 464–470.

Pettersson, A., et al. (2012), The TMPRSS2:ERG rearrangement, ERG expression, and prostate cancer outcomes: a cohort study and meta-analysis, *Cancer Epidemiology, Biomarkers & Prevention*, *21*(9), 1497–1509.

Rabiner, L. R. (1989), A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, *77*(2), 257–286.

Ries, L. A. G., J. L. Young Jr, G. E. Keel, M. P. Eisner, Y. D. Lin, and M.-J. D. Horner (2007), Cancer survival among adults: US SEER program, 1988–2001, *Patient and tumor characteristics SEER Survival Monograph Publication*, pp. 07–6215.

Risko, R., S. Merdan, P. R. Womble, C. Barnett, Z. Ye, S. M. Linsell, J. E. Montie, D. C. Miller, and B. T. Denton (2014), Clinical predictors and recommendations for staging computed tomography scan among men with prostate cancer, *Urology*, *84*(6), 1329–1334.

Roehl, K. A., M. Han, C. G. Ramos, J. A. V. Antenor, and W. J. Catalona (2004), Cancer progression and survival rates following anatomical radical retropubic prostatectomy in 3,478 consecutive patients: long-term results, *The Journal of Urology*, *172*(3), 910–914.

Ross, K. S., H. B. Carter, J. D. Pearson, and H. A. Guess (2000), Comparative efficiency of prostate-specific antigen screening strategies for prostate cancer detection, *Journal of the American Medical Association*, *284*(11), 1399–1405.

Roth, J., R. Gulati, J. Gore, M. Cooperberg, and R. Etzioni (2016), Economic analysis of prostate-specific antigen screening and selective treatment strategies, *JAMA Oncology*, *2*(7), 890–898.

Salagierski, M., and J. A. Schalken (2012), Molecular diagnosis of prostate cancer: PCA3 and TMPRSS2:ERG gene fusion, *The Journal of Urology*, *187*(3), 795–801.

Salami, S. S., et al. (2013), Combining urinary detection of TMPRSS2:ERG and PCA3 with serum PSA to predict diagnosis of prostate cancer, in *Urologic Oncology: Seminars and Original Investigations*, vol. 31, pp. 566–571, Elsevier.

Sartori, D. A., and D. W. Chan (2014), Biomarkers in prostate cancer: whats new?, *Current Opinion in Oncology*, *26*(3), 259.

Savage, C. J., H. Lilja, A. M. Cronin, D. Ulmert, and A. J. Vickers (2010), Empirical estimates of the lead time distribution for prostate cancer based on two independent representative cohorts of men not subject to prostate-specific antigen screening, *Cancer Epidemiology, Biomarkers & Prevention*, *19*(5), 1201–1207.

Schröder, F. H., et al. (2009), Screening and prostate-cancer mortality in a randomized European study, *New England Journal of Medicine*, *360*(13), 1320–1328.

Schröder, F. H., et al. (2012), Prostate-cancer mortality at 11 years of follow-up, *New England Journal of Medicine*, *366*(11), 981–990.

Schröder, F. H., et al. (2014), Screening and prostate cancer mortality: results of the European randomised study of screening for prostate cancer (ERSPC) at 13 years of follow-up, *The Lancet*, *384*(9959), 2027–2035.

Shepard, D. (1996), *Cost-effectiveness in Health and Medicine. Edited by MR Gold, JE Siegel, LB Russell, and MC Weinstein*, 1 ed., Oxford University Press, New York, New York.

Shi, J., O. Alagoz, F. S. Erenay, and Q. Su (2014), A survey of optimization models on cancer chemotherapy treatment planning, *Annals of Operations Research*, *221*(1), 331–356.

Siddiqui, M. M., et al. (2015), Comparison of MR/ultrasound fusion–guided biopsy with ultrasound-guided biopsy for the diagnosis of prostate cancer, *Journal of the American Medical Association*, *313*(4), 390–397.

Siddiqui, M. M., et al. (2016), Efficiency of prostate cancer diagnosis by MR/ultrasound fusion-guided biopsy vs standard extended-sextant biopsy for MR-visible lesions, *Journal of the National Cancer Institute*, *108*(9), djw039.

Simmons Ivy, J., H. Black Nembhard, and K. Baran (2009), Quantifying the impact of variability and noise on patient outcomes in breast cancer decision making, *Quality Engineering*, *21*(3), 319–334.

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

Tejada, J. J., J. S. Ivy, R. E. King, J. R. Wilson, M. J. Ballan, M. G. Kay, K. M. Diehl, and B. C. Yankaskas (2014), Combined DES/SD model of breast cancer screening for older women, II: screening-and-treatment simulation, *IIE Transactions*, *46*(7), 707–727.

Tejada, J. J., J. S. Ivy, J. R. Wilson, M. J. Ballan, K. M. Diehl, and B. C. Yankaskas (2015), Combined DES/SD model of breast cancer screening for older women, I: Natural-history simulation, *IIE Transactions*, *47*(6), 600–619.

Terris, M. K. (1999), Sensitivity and specificity of sextant biopsies in the detection of prostate cancer: preliminary report, *Urology*, *54*(3), 486–489.

Thompson, I. M., D. P. Ankerst, C. Chi, P. J. Goodman, C. M. Tangen, M. S. Lucia, Z. Feng, H. L. Parnes, and C. A. Coltman (2006), Assessing prostate cancer risk: results from the prostate cancer prevention trial, *Journal of the National Cancer Institute*, *98*(8), 529–534.

Tomlins, S. A., et al. (2005), Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer, *Science*, *310*(5748), 644–648.

Tomlins, S. A., et al. (2016), Urine TMPRSS2:ERG plus PCA3 for individualized prostate cancer risk assessment, *European Urology*, *70*(1), 45–53.

Tosoian, J. J., B. J. Trock, P. Landis, Z. Feng, J. I. Epstein, A. W. Partin, P. C. Walsh, and H. B. Carter (2011), Active surveillance program for prostate cancer: an update of the Johns Hopkins experience, *Journal of Clinical Oncology*, *29*(16), 2185–2190.

Tosoian, J. J., A. E. Ross, L. J. Sokoll, A. W. Partin, and C. P. Pavlovich (2016), Urinary biomarkers for prostate cancer, *Urologic Clinics of North America*, *43*(1), 17–38.

Truong, M., B. Yang, and D. F. Jarrard (2013), Toward the detection of prostate cancer in urine: a critical analysis, *The Journal of Urology*, *189*(2), 422–429.

Tsodikov, A., A. Szabo, and J. Wegelin (2006), A population model of prostate cancer incidence, *Statistics in Medicine*, *25*(16), 2846–2866.

Underwood, D., J. Zhang, B. Denton, N. Shah, and B. Inman (2012), Simulation optimization of PSA-threshold based prostate cancer screening policies, *Health Care Management Science*, *15*(4), 293–309.

van den Bergh, R. C., S. Roemeling, M. J. Roobol, W. Roobol, F. H. Schröder, C. H. Bangma, et al. (2007), Prospective validation of active surveillance in prostate cancer: the PRIAS study, *European Urology*, *52*(6), 1560–1563.

Wade, J., et al. (2013), Psychological impact of prostate biopsy: Physical symptoms, anxiety, and depression, *Journal of Clinical Oncology*, *31*(33), 4235–4241.

Willis, S. R., H. U. Ahmed, C. M. Moore, I. Donaldson, M. Emberton, A. H. Miners, and J. van der Meulen (2014), Multiparametric MRI followed by targeted prostate biopsy for men with suspected prostate cancer: a clinical decision analysis, *BMJ Open*, *4*(6), e004,895.

Young, A., N. Palanisamy, J. Siddiqui, D. Wood, J. Wei, A. Chinnaiyan, L. Kunju, and S. Tomlins (2012), Correlation of urine TMPRSS2:ERG and PCA3 to ERG+ and total prostate cancer burden, *American Journal of Clinical Pathology*, *138*(5), 685–696.

Zhang, J. (2011), Partially observable markov decision processes for prostate cancer screening, Ph.D. thesis, North Carolina State University.

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

Zhang, J., B. T. Denton, H. Balasubramanian, N. D. Shah, and B. A. Inman (2012b), Optimization of PSA screening policies: a comparison of the patient and societal perspectives, *Medical Decision Making*, *32*(2), 337–349.

Zhang, Y., B. T. Denton, and M. E. Nielsen (2013), Comparison of surveillance strategies for low-risk bladder cancer patients, *Medical Decision Making*, *33*(2), 198–214.