Optimization and Machine Learning Methods for Diagnostic Testing of Prostate Cancer

by

Selin Merdan

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
2018

Doctoral Committee:

Professor Brian T. Denton, Chair
Professor Mark Daskin
Associate Professor Henry Lam
Assistant Professor Jenna Wiens

Selin Merdan
smerdan@umich.edu
ORCID iD: 0000-0001-9226-959X

Dedication

This dissertation is dedicated to Petter, my parents, brother, aunts and grandparents.
Thank you for your unconditional love and support.
I love you dearly.

Acknowledgments

First, I would like to thank my advisor, Professor Brian Denton, for his immense guidance and support over the past six 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. Any amount of success I may have, either now or in the future, also belongs to him. Furthermore, I gratefully acknowledge the support that I received from the National Science Foundation with grant number CMMI 1536444, the Rackham Predoctoral Fellowship, and the Seth Bonder Departmental Fellowship at the University of Michigan.

I would like thank my committee members Professor Mark Daskin, Professor Henry Lam, and Professor Jenna Wiens for serving on my committee and providing me with career advice and helpful feedback on my research. I also would like to acknowledge Professor Jon Lee who inspired me to pursue graduate research and to become an avid cyclist. In addition, I would like to thank our collaborators from Michigan Medicine, Dr. James Montie, Dr. David Miller, Dr. Gregory Auffenberg, Dr. Todd Morgan, Dr. Scott Tomlins, and Dr. John Wei for their invaluable clinical perspective and for teaching me about the challenges in early detection of prostate cancer. In particular, I owe much gratitude to Dr. David Miller and Dr. James Montie. My summer internship under your mentorship constituted the major driving force behind my decision to pursue Ph.D. degree with an emphasis on medical decision-making.

Thank you to my classmates, officemates and many friends in the department and beyond, too many to list all by name, for your support. Finally, I would like to thank all of my family who have continuously encouraged and supported me, and without whom I would not be at this stage of my graduate school career. Thank you to my parents, my brother and my aunts for their unwavering love and support throughout my life. Most of all, thank you Petter for your love, wisdom, perspective, and for always believing in me.

Table of Contents

List of Figures

List of Tables

List of Appendices

List of Acronyms

ACS American Cancer Society.

ADT androgen deprivation therapy.

AUA American Urology Association.

AUC area under the ROC curve.

BS bone scan.

CCI Charlson comorbidity index.

CI confidence interval.

Cos-LapKLR cost-sensitive Laplacian kernel logistic regression.

Cos-LR cost-sensitive logistic regression.

Cos-SVM cost-sensitive support vector machines.

CT scan computed tomography.

CV cross validation.

DRE digital rectal examination.

EAU European Association of Urology.

GS Gleason score.

KLR kernel logistic regression.

LP linear program.

LR logistic regression.

MCKP multiple choice knapsack problem.

MIP mixed integer program.

MLE maximum likelihood estimation.

MRI magnetic resonance imaging.

MUSIC Michigan Urological Surgery Improvement Collaborative.

NCCN National Comprehensive Cancer Network.

PCa prostate cancer.

PCA3 prostate cancer antigen 3 assay.

PET positron emission tomography.

PSA prostate-specific antigen.

RF Random forests.

ROC receiver operating characteristics.

ROS random oversampling.

RP radical prostatectomy.

RT radiation therapy.

RUS random undersampling.

SVM support vector machines.

T2:ERG TMPRSS2:ERG assay.

Abstract

Technological advances in biomarkers and imaging tests are creating new avenues to advance precision health for early detection of cancer. These advances have resulted in multiple layers of information that can be used to make clinical decisions, but how to best use these multiple sources of information is a challenging engineering problem due to the high cost and imperfect sensitivity and specificity of these tests. Questions that need to be addressed include which diagnostic tests to choose and how to best integrate them, in order to optimally balance the competing goals of early disease detection and minimal cost and harm from unnecessary testing. To study these research questions, we present new optimization-based models and data-driven analytic methods in three parts to improve early detection of prostate cancer (PCa).

In the first part, we develop and validate predictive models to assess individual PCa risk using known clinical risk factors. Because not all men with newly-diagnosed PCa received imaging at diagnosis, we use an established method to correct for verification bias to evaluate the accuracy of published imaging guidelines. In addition to the published guidelines, we implement advanced classification modeling techniques to develop accurate classification rules identifying which patients should receive imaging. We propose a new algorithm for a classification model that considers information of patients with unverified disease and the high cost of misclassifying a metastatic patient. We summarize our development and implementation of state-wide, evidence-based imaging criteria that weigh the benefits and harms of radiological imaging for detection of metastatic PCa.

In the second part of this thesis, we combine optimization and machine learning approaches into a robust optimization framework to design imaging guidelines that can account for imperfect calibration of predictions. We investigate efficient and effective ways to combine multiple medical diagnostic tests where the result of one test may be used to predict the outcome of another. We analyze the properties of the proposed optimization models from the perspectives of multiple stakeholders, and we present the results of fast

approximation methods that we show can be used to solve large-scale models.

In the third and final part of this thesis, we investigate the optimal design of *composite* multi-biomarker tests to achieve early detection of prostate cancer. Biomarker tests vary significantly in cost, and cause false positive and false negative results, leading to serious health implications for patients. Since no single biomarker on its own is considered satisfactory, we utilize simulation and statistical methods to develop the optimal diagnosis procedure for early detection of PCa consisting of a sequence of biomarker tests, balancing the benefits of early detection, such as increased survival, with the harms of testing, such as unnecessary prostate biopsies.

In this dissertation, we identify new principles and methods to guide the design of early detection protocols for PCa using new diagnostic technologies. We provide important clinical evidence that can be used to improve health outcomes of patients while reducing wasteful application of diagnostic tests to patients for whom they are not effective. Moreover, some of the findings of this dissertation have been implemented directly into clinical practice in the state of Michigan. The models and methodologies we present in this thesis are not limited to PCa, and can be applied to a broad range of chronic diseases for which diagnostic tests are available.

Chapter 1.

Introduction

Early detection of cancer can lower the incidence rate and prolong survival by detecting disease at early stages when treatment outcomes are most favorable for patients. Recent advances in diagnostic technology, including genomics, biomarkers and radiological imaging offer the potential for early detection of cancer. However, these advances have made clinical decision making difficult because the tests are not sufficiently *sensitive* and *specific* on their own. Sensitivity is the probability that a test is positive given the disease is present, and specificity is the probability that the test is negative given the disease is not present. Moreover, there is often a tradeoff between the benefits of an accurate diagnosis of the anticipated disease and harms and costs associated with the diagnostic tests themselves. It is therefore challenging to determine how to use tests optimally.

Prostate cancer (PCa) is the perfect test-bed to explore these challenging problems because of (1) its societal importance as the most common cancer among American men; and (2) many new diagnostic tests have been discovered; however, it is unclear how to best utilize them. PCa is now the second most commonly diagnosed cancer (more than $160,000$ new cases are expected in 2018) and is also the second leading cause of death from cancer among American men (more than $25,000$ deaths estimated in 2018) [2]. The chance of a man being diagnosed with PCa in his lifetime is 1 in 9 [2]. The management of PCa is challenging because there are multiple types of cancer, ranging from likely indolent to likely lethal. As a result, there is a need for individualized strategies that can judiciously identify men in need of diagnosis and limit the costly and invasive nature of diagnostic testing for those men who will not benefit. To accurately predict the prognosis and ensure appropriate treatment of patients with PCa, accurate diagnosis and staging are crucial.

The risk of developing PCa varies among patients depending on many factors such as

advanced age, ethnicity, and family history of the disease. More than 65% of all PCa cases occur in men older than 65 years [3]. African American men have the highest incidence of PCa in the world. The current diagnosis of PCa is made based on risk stratification by the combination of digital rectal examination (DRE) and serum prostate-specific antigen (PSA) level. Despite its widespread use at diagnosis, establishing a PSA cutoff that can reliably indicate the presence of cancer or the need for a biopsy is not possible due to the poor sensitivity and specificity of serum PSA [11, 134]. Because serum PSA is a gland-specific rather than cancer-specific biomarker, it does not reliably distinguish either cancer from benign prostatic conditions, or clinically significant from likely indolent cancers.

In recent years, several molecular biomarkers and corresponding diagnostic assays have been developed with the potential to improve early detection of PCa. In particular, urine biomarkers are attractive because they are noninvasive and can be used for prediction purposes. They are based on DNA, RNA or protein analysis in urine. Prostate cancer antigen 3 assay (PCA3) and TMPRSS2:ERG assay (T2:ERG) gene fusions are the most advanced PCa-specific early detection biomarkers that have been shown to predict biopsy outcome more accurately than PSA, and reduce the likelihood of false positive results [34, 123, 137, 139, 168, 179].

Patients who have suspicious clinical findings and biomarker test results are further evaluated with biopsy, which is currently the gold standard test to confirm PCa. 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 [158]. Biopsies expose patients to additional anxiety and complications (e.g., bleeding, urinary retention and sepsis), and the risks of a diagnosis of an indolent PCa which can lead to invasive procedures and treatments [172]. If cancer cells are found upon evaluation of the biopsy by a pathologist, the cells are given a Gleason score (GS). 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 GS between 2 and 10. A higher GS indicates that the tumor is more likely to grow and spread quickly.

Because PSA is not cancer-specific, many men undergo one or more repeat biopsies after an initial negative biopsy [37, 143]. Men with prior negative biopsies, but persistent suspicion of PCa (e.g., a persistent elevated/rising PSA level and/or a suspicious DRE) pose a diagnostic challenge. The positive predictive values of DRE and serum PSA are

relatively low, such that only 1 in 4 repeat biopsies will reveal PCa [164]. PCA3 and T2:ERG have also been shown to better predict repeat biopsy outcomes in men with elevated serum PSA levels and previous negative biopsy findings. Although there are studies supporting increased diagnostic accuracy for both of these biomarkers, no previous study has compared these biomarkers to determine the ideal thresholds to trigger a repeat biopsy, and the resulting increase in survival and decrease in unnecessary biopsies.

Once a PCa diagnosis has been made, the urologist works to determine the extent (stage) of the cancer. The most significant health outcome to consider when determining the cancer stage is whether the cancer has metastasized (i.e., spread to other parts of the body). Metastases are associated with significant morbidity, increased mortality and substantial economic burden [31, 178]. Staging PCa is important not only for prediction of prognosis, but also for choosing the optimal course of treatment. During staging, the urologist may order a bone scan (BS) and/or a computed tomography (CT scan) to detect bone and lymph node metastases, respectively, which are the most commonly used imaging tests. However, not all patients with PCa have the same risk of harboring metastatic disease at diagnosis, thus not every patient should have every test.

There are harms associated with both over-imaging and missing a patient with metastasis. One of the risks associated with certain types of imaging, such as CT scan, is that they expose patients to potentially harmful radiation. The effects of radiation add up over a patient's lifetime. A study by Smith-Bindman et al. found the median effective dose of abdomen-pelvis CT scan to be 32 millisievert (mSv), and it was concluded that at that dose 1 in every 660 60-year-old men receiving an abdomen-pelvis CT scan will develop cancer from the procedure [149]. Moreover, according to the Life Span Study cohort of atomic bomb survivors, exposure to 32 mSv significantly increases the relative risk for developing cancer [124]. Imaging tests can also result in false positives that lead to stress, more tests, and treatment that is unlikely to benefit the patient. At the same time, these studies are expensive and time-consuming, and the overall yield (i.e., the likelihood of detecting metastases) is quite low for men with low- or intermediate-risk cancers.

Multiple clinical guidelines indicating the need for imaging in patients with certain risk factors have been established; however, there is no consensus regarding the optimal use of staging BS and CT scan for men with newly-diagnosed PCa. This causes persistent variation in the use of these tests in practice, including potentially unnecessary testing in many men at low risk for metastatic disease and the absence of testing for some men with higher-risk cancers. Underscoring the significance of this issue, the American Urology

Association (AUA) recently identified the avoidance radiological imaging in men with low-risk prostate cancer as its number one priority for the national *Choosing Wisely* program [6].

In determining which diagnostic tests should be used in the evaluation of metastases in PCa, it is important to recognize the strengths and limitations of each test. Currently, imaging for metastatic disease involves the application of BS and CT scan. The implications of this approach include patient time, imaging time, costs and radiation exposure. However, there is no clinical guideline addressing the need for both BS and CT scan. The correlation between imaging test results motivates a sequential testing paradigm in which some patients may benefit from having tests one at a time so that the results of one test can be used to predict the outcome of the follow-on test, with the potential to use the individual diagnostic resources more efficiently and effectively.

Summary of contributions. Chapter 2 uses a novel collection of methods including statistics, machine learning, and optimization methods, that we collectively refer to as *data-analytics* methods, to determine which patients should receive a staging BS and/or a CT scan and which patients can safely avoid imaging on the basis of individual risk factors. The main contributions in the chapter are as follows:

- *Risk Prediction Models for Metastatic Prostate Cancer.* We develop new risk prediction models that accurately estimate the probability of a positive imaging test. We perform internal validation of these models via bootstrapping and an out-of-sample evaluation of the predictions. These models are subsequently used to evaluate the diagnostic accuracy of imaging guidelines accounting for the bias introduced by the patients with nonverified disease status, and to optimize imaging guidelines for which patients should receive a BS or CT scan.

- *Classification Modeling for Metastatic Cancer Detection.* We utilize existing optimization and machine learning methods, and compare these to a new approach we propose to design classification rules that distinguish metastatic patients from patients with localized cancer. To our knowledge, this is the first study to employ classification modeling techniques in the detection of cancer considering (1) the exploitation of data for the patients who did not have the gold standard tests (either BS or CT scan) at diagnosis and (2) the incorporation of a cost-sensitive learning scheme to deal with the class imbalance problem simultaneously in the learning framework.

- *Bias-corrected Performance of Imaging Guidelines.* Because not all men with newly-diagnosed prostate cancer underwent imaging, we apply statistical methods to mitigate verification bias to evaluate the diagnostic accuracy of imaging guidelines for detection of metastatic disease. Our definition of imaging guidelines is the union of previously published clinical guidelines and optimized classification rules we develop using machine learning methods.

- *Implementation and Measurement of Impact.* Following adoption of the guidelines, the impact on BS and CT scan utilization was evaluated to confirm the predicted results that indicated a similar or improved detection rate and substantial reductions in unnecessary imaging. Therefore, this chapter also serves as a case study of the practical implementation of data-analytics methods with measurable impact.

Chapter 3 investigates the optimal design of coordinated imaging protocols considering different combinations of imaging tests to improve detection of metastatic cancer while at the same time reducing unnecessary testing. The main contributions of the chapter are as follows:

- *Robust Coordinated Imaging Protocols for Metastatic Cancer Detection.* To our knowledge, we are the first to integrate robust optimization models with predictive models to optimize diagnostic testing decisions pertaining to the selection of imaging protocols for patient population. Furthermore, we propose models for sequential testing where the outcome of one test informs the decision about the follow-up test. These models are used to address the lack of a standardized holistic approach for recommending imaging tests on the basis of individuals' risk of disease while accounting for errors in predictions.

- *Clinically Acceptable Heuristics.* We propose heuristics that incorporate the perspectives of multiple stakeholders participating in the decision making process for imaging and that lead to more predictable decisions than solving an optimization model. We evaluate the worst-case and average case behavior of the proposed heuristics using test cases based on real data.

- *Case Study on Medical Data.* We use medical data from a large statewide collaborative to answer important questions regarding the benefits of multi-modality imaging for PCa staging.

Chapter 4 presents decision analysis for men with elevated PSA to evaluate the value of PCA3 and T2:ERG in the diagnosis of PCa at repeat biopsy by comparing the loss in the overall survival to the gain in repeat biopsy rate. The main contributions of the chapter are as follows:

- *Head-to-head Comparison of Biomarkers.* We are the first to investigate the long-term health outcomes associated with the use of two new and promising biomarkers, PCA3 and T2:ERG, in men with at least one previous negative biopsy and elevated serum PSA when making repeat biopsy decisions for clinically localized PCa (i.e., cancer confined in prostate). We present a decision analysis model and results of sensitivity analysis to assess the impact of model parameter uncertainty using Monte Carlo simulation.

- *Clinical and Policy Implications.* We use our decision model to answer key questions about early detection of PCa, such as whether and how to use newly discovered biomarkers effectively to better select men for repeat biopsy. We show that the use of PCA3 testing or T2:ERG for repeat biopsy decisions can reduce the number of biopsies substantially without significantly affecting survival.

The remainder of this thesis presents the work described above in Chapters 2 - 4. The thesis is concluded in Chapter 5 with a summary of the most important findings and an outline of future research opportunities.

Chapter 2.

Optimal Design of Imaging Guidelines for Detection of Metastatic Prostate Cancer

2.1. Introduction

Prostate cancer (PCa) is the most common cancer among men. It has been estimated that in 2017 there will be more than 160,000 new cases of PCa diagnosed in the United States. For each of these cases, clinical *staging* will be performed to determine the extent of the disease. The most significant health outcome to consider when determining the stage of PCa is whether the cancer has metastasized (i.e., spread to other parts of the body), since this will determine the optimal course of treatment. During staging, the urologist may order a bone scan (BS) and/or a computed tomography (CT scan), because they are the most frequently used noninvasive imaging methods to detect bone and lymph node metastases, respectively.

Optimal treatment of men with newly-diagnosed PCa depends on the stage of disease at diagnosis. Accordingly, the performance of a staging BS and CT scan is pivotal to the diagnostic evaluation and treatment planning for some men with PCa. At the same time, however, there are harms associated with both over- and under-imaging. Under-imaging results in patients' metastatic PCa going undetected. In such cases, patients are subjected to treatment, such as radical prostatectomy (surgical removal of the prostate), that is unlikely to benefit the patient, and can lead to serious side effects and negative health outcomes due to delays in chemotherapy. Over-imaging causes potentially harmful

radiation exposure and often results in incidental findings that require follow-up procedures that can be painful and risky for the patient. Additionally, unnecessary imaging blocks access to the imaging resources for other patients, and unnecessarily increases healthcare costs.

There are several international evidence-based guidelines indicating the need for BS and CT scan only in patients with certain unfavorable risk factors (see Table 2.7); however, the guidelines vary in their recommendations and there is no consensus about the optimal use of BS and CT scan for men newly diagnosed with PCa [30, 36, 74, 115, 117, 161]. The net effect is that imaging practice patterns continue to vary widely, implying immediate opportunities to improve value in this area of PCa care. Many believe that an important next step in this process is to move away from recommendations based on the risk of recurrence after treatment (e.g., D'Amico risk groups), and toward the identification and implementation of imaging criteria that most accurately forecast a positive study that would actually change clinical decision-making. To address this issue, we took a holistic perspective to determine which patients should receive a BS and/or a CT scan and which patients can safely avoid imaging on the basis of individual risk factors. We evaluated our proposed data-driven approaches in a population-based sample of men with newly-diagnosed prostate cancer from the diverse academic and community practices in the Michigan Urological Surgery Improvement Collaborative (MUSIC), which includes 90% of the urologists in the state of Michigan (see `http://musicurology.com/`).

Figure 2.1 illustrates the linkages between each of the components of the research design for this project from data processing to implementation. The remainder of this chapter is structured as follows. Section 2.2 describes the methodological approach for development and validation of risk prediction models, and proper measures for evaluating prediction performance. Section 2.3 reviews the challenges of classification modeling in imbalanced observational health data and describes our proposed algorithm for cost-sensitive semi-supervised learning. Section 2.4 provides background on the problem of verification bias and describes the methodological approach we considered in tackling the bias for correcting the diagnostic accuracy of imaging guidelines. Section 2.5 describes the implementation process and the impact of our work based on post-implementation analysis. Section 2.6 highlights our main conclusions and states some points for future research.

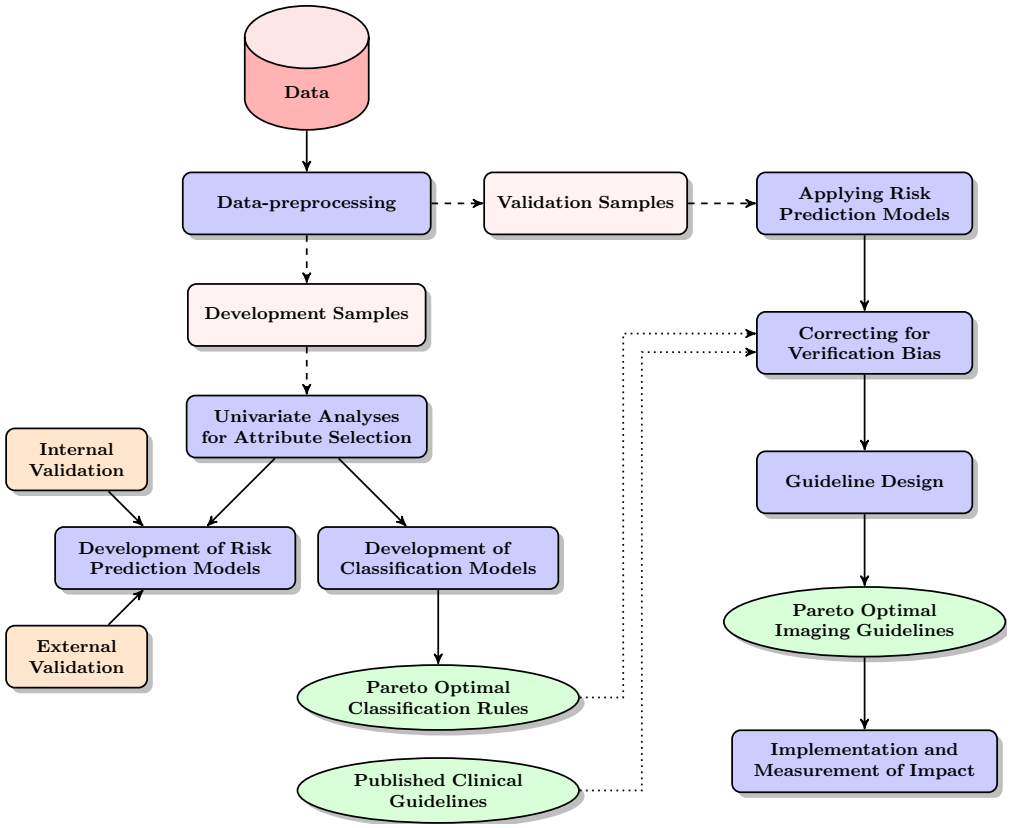

Figure 2.1: Research framework illustrating the major steps from data-preprocessing to implementation and measurement of impact.

2.2. Risk Prediction Models for Metastatic Prostate Cancer

For a risk prediction model to be useful for personalized medicine and patient counseling, it is necessary to ensure the model is calibrated to provide reliable predictions for the patients. This section describes the development and testing of predictive models for estimating the probability of an imaging test that was positive for metastases.

2.2.1. Clinical Datasets and Variables

Established in 2011 with funding from Blue Cross Blue Shield of Michigan, MUSIC is a consortium of 43 practices from throughout Michigan that aims to improve the quality and cost-efficiency of care provided to men with prostate cancer. Each practice involved in MUSIC obtained an exemption or approval for participation from a local institutional

review board.

PCa is diagnosed by biopsy, which involves extraction of tissue (normally 12 samples) from the prostate. These samples produce useful predictors of metastasis, such as a pathology grading called Gleason score (GS), percentage of positive samples (also called *cores*) that show cancer, and the maximum percent core involvement. These risk factors are determined by review of biopsy samples by a trained pathologist. GS is a pathological characterization of the cancer cells that is correlated with the risk of metastasis, and the percentage of positive cores and the maximum core involvement is correlated with tumor volume. Other potentially relevant risk factors for metastasis include a patient's age, prostate-specific antigen (PSA) score, and clinical T stage. A PSA test is a simple blood test that indicates the amount of PSA, a protein produced by cells of the prostate gland, that escapes into the blood from the prostate. Patients with higher than normal PSA values have a greater risk of metastatic PCa. Clinical T stage is part of the TNM staging system for PCa that defines the extent of the primary tumor based on clinical examination.

The MUSIC registry contains detailed clinical and demographic information, including patient age, serum PSA at diagnosis, clinical T stage, biopsy GS, total number of biopsy cores, number of positive cores, and the receipt and results of imaging tests ordered by the treating urologist. The initial analysis for BS included $1,519$ patients with newly-diagnosed PCa seen at 19 MUSIC practices in Michigan from March 2012 through June 2013. Among this group, $416\,(27.39\%)$ patients underwent staging BS. Among the patients that received a BS, $48\,(11.54\%)$ had a positive outcome with evidence for bone metastasis. The cohort for CT scan included $2,380$ men with newly diagnosed PCa from 27 MUSIC practices from March 2012 to September 2013. Among $2,380$ patients, $643\,(27.02\%)$ of them underwent a staging CT scan, and $62\,(9.64\%)$ of these studies were interpreted as positive for metastasis.

As a first step, we compared clinical and pathological characteristics of patients with or without imaging. Differences between these two groups of patients in medians for quantitative variables, and differences in distributions for categorical variables, were compared using Mann-Whitney's U-test, and Chi-square test, respectively. We performed univariate and multivariate analyses to examine the association between imaging outcomes and all routinely available clinical variables in imaged patients.

Table 2.1 presents clinical characteristics of the $1,509$ patients with newly-diagnosed PCa. Patients who received staging BS had higher mean PSA values as well as higher percentages of positive cores compared to patients without BS (all $p \leq 0.001$). Moreover,

Table 2.1: Patient characteristics for BS cohort.

Variables	All patients without BS ($\mathbf{n = 1,103}$)	All patients with BS ($\mathbf{n = 416}$)	p-value
Age, (years)			0.02
Mean (median)	64.2 (64.4)	68.2 (67.7)	
Range	$40.4 - 95.8$	$41.8 - 90.5$	
Clinical stage, No. (%)			< 0.0001
T1	881 (79.9)	216 (51.9)	
T2	214 (19.4)	173 (41.6)	
T3/4	8(0.7)	27 (6.5)	
PSA, ng/mL			0.003
Mean (median)	8.0 (5.2)	61.8 (7.7)	
Range	$0.2 - 620.8$	$0.4 - 6873.4$	
PSA, ng/mL, No. (%)			< 0.0001
≤ 10	1018 (92.3)	247 (59.4)	
$10.1 - 20$	58 (5.3)	81 (19.5)	
$20.1 - 50$	10 (0.9)	45 (10.8)	
$50.1 - 100$	12 (1.1)	20 (4.8)	
> 100	5 (0.5)	23 (5.5)	
Biopsy Gleason sum, No. (%)			< 0.0001
≤ 6	488 (44.2)	33 (7.9)	
$3 + 4$	439 (39.8)	105 (25.2)	
$4 + 3$	137 (12.4)	58 (13.9)	
$8 - 10$	39 (3.6)	220 (52.9)	
Biopsy cores taken, No.			0.50
Mean (median)	12.5 (12.0)	12.9 (12.0)	
Range	$4 - 82$	$1 - 78$	
Positive cores, No.			0.0004
Mean (median)	3.2 (3.0)	6.3 (6.0)	
Range	$0 - 20$	$1 - 16$	
Positive cores, %			< 0.0001
Mean (median)	26.4 (21.1)	51.2 (50.0)	
Range	$0 - 100$	$3.1 - 100$	

patients with BS were significantly older and showed a higher GS as well as higher rate of locally advanced PCa compared to patients without BS (all $p \leq 0.001$). Table 2.2 summarizes results from univariate and multivariate analyses evaluating the relationship between clinical parameters and BS findings. There was a wide range of serum PSA values, ($0.4-6873.4$ ng/mL, coefficient of variation 651.2), and due to the dispersion in PSA levels, we used the natural logarithm transformation. In univariate logistic regression analyses, all variables were significant predictors of bone metastases (all $p \leq 0.01$). In multivariable analyses, only serum PSA and biopsy GS were significant predictors of a positive BS (both p-values ≤ 0.004) (Table 2.2). Illustrating this point, the adjusted odds of a positive BS for

patients with a biopsy GS $4+3=7$ are 3.30 (95% confidence interval (CI): $0.55-19.89$) times as great as for patients with GS $3+4=7$ or GS $=6$, while for patients with biopsy GS $8-10$, the odds of a positive BS are 9.53 (95% CI: $2.14-42.38$) times the odds for patients in the reference group.

Table 2.2: Univariable and multivariable logistic regression (LR) models predicting the presence of bone metastases at diagnosis.

	Univariable logistic regression model		**Multivariable logistic regression model**		
Variables	**OR (95% CI)**	**p-value**	**OR (95% CI)**	**p-value**	**Overall p-value**
Age at diagnosis (year)	1.04 (1.01 – 1.08)	0.01	1.03 (0.99 – 1.06)	0.14	(0.14)
ln(PSA+1), ng/mL	2.25 (1.76 – 2.88)	< 0.0001	2.00 (1.51 – 2.64)	< 0.0001	(< 0.0001)
Biopsy Gleason score, No. (%)					(0.004)
$\leq 3+4$	Reference		Reference		
$4+3$	5.04 (0.90 – 28.31)	0.07	3.30 (0.55 – 19.89)	0.19	
$8-10$	16.05 (3.82 – 67.45)	0.0002	9.53 (2.14 – 42.38)	0.003	
Clinical T stage, No. (%)					(0.4)
T1	Reference		Reference		
T2	2.64 (1.31 – 5.33)	0.007	1.61 (0.72 – 3.57)	0.25	
T3/4	9.19 (3.51 – 24.03)	< 0.0001	1.91 (0.57 – 6.43)	0.30	
Positive cores, %	13.32 (4.26 – 41.72)	< 0.0001	1.70 (0.42 – 6.90)	0.46	(0.46)

Table 2.3 presents the clinical characteristics of 2,380 patients included in the analytic sample. Patients who underwent CT scan imaging had significantly higher PSA levels, biopsy GS, and clinical T stages than those who did not receive a CT scan scan (all $p < 0.0001$). Table 2.4 summarizes results from the univariate and multivariate logistic regression models, and presents the associations between clinical variables and a positive CT scan scan. The univariate analyses identified PSA, GS, clinical stage, and the ratio of positive cores as statistically significant predictors of a positive study (all p-values < 0.0001). In the multivariate analysis, PSA, GS ≥ 8, and clinical stage $\geq$ T3 were predictors of metastases (all $p < 0.05$) (Table 2.4). A separate model with PSA as a categorical variable revealed that PSA > 20 was a statistically significant cutoff. Illustrating this point, for the multivariate logistic regression model the odds ratio for PSA in the range 10.1 to 20 was 1.92 (95%CI : $0.82-4.49$), compared to 5.37 (95%CI : $2.52-11.44$) for PSA > 20.

We included all variables with a statistically significant association which were as follows: age at diagnosis, natural logarithm of PSA+1 (ln(PSA+1)), biopsy GS ($\leq 3+4$, $4+3$, or $8-10$), clinical T stage (T1, T2, or T3/4) and the percentage of positive biopsy cores.

Table 2.3: Patient characteristics for CT scan cohort.

Variables	All patients without CT ($\mathbf{n = 1,737}$)	All patients with CT ($\mathbf{n = 643}$)	p-value
Age, (years)			0.17
Mean (median)	63.8 (64)	66.0 (66)	
Range	$40.4 - 95$	$40 - 99$	
Clinical stage, No. (%)			< 0.0001
T1	1,386 (79.8)	359 (55.8)	
T2	339 (19.5)	246 (38.3)	
T3/4	12 (0.69)	38 (5.91)	
PSA, ng/mL			< 0.0001
Mean (median)	8.6 (5.2)	49.9 (7.7)	
Range	0.21, 008.9	$0.40 - 6,873.40$	
PSA, ng/mL, No. (%)			< 0.0001
≤ 10	1576 (90.7)	377 (58.6)	
$10.1 - 20$	124 (7.1)	146 (22.7)	
$20.1 - 50$	20 (1.2)	64 (10.0)	
> 50	17 (1.0)	56 (8.7)	
Biopsy Gleason sum, No. (%)			< 0.0001
≤ 6	747 (43.0)	62(9.6)	
$3 + 4$	671 (38.6)	174 (27.1)	
$4 + 3$	212 (12.2)	97 (15.1)	
$8 - 10$	107 (6.2)	310 (48.2)	
Biopsy cores taken, No.			0.4
Mean (median)	12.5 (12.0)	12.7 (12.0)	
Range	$2 - 82$	$1 - 78$	
Positive cores, No.			< 0.0001
Mean (median)	3.3 (3.0)	6.2 (6.0)	
Range	$1 - 20$ $1 - 16$		
Positive cores, %			< 0.0001
Mean (median)	27.0 (23.1)	50.4 (50.0)	
Range	$2.4 - 100$	$3.1 - 100$	

2.2.2. Predictive Models

Suppose that l patients have been imaged and we are given the empirical training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l) \in \mathbb{R}^d \times \{\pm 1\}$ of those patients, where y_i's are the binary imaging outcomes and d is the number of patient attributes (e.g., age, GS, PSA, etc.). Let $\mathbf{X} \in \mathbb{R}^{l \times d}$ be the data matrix and $\mathbf{y}$ be the binary vector of imaging outcomes. For every attribute vector $\mathbf{x}_i \in \mathbb{R}^d$ (a row vector in $\mathbf{X}$), where $i = 1, \ldots, l$, the outcome is either $y_i = 1$ or $y_i = -1$; where 1 corresponds to a positive test and -1 to a negative test. We assume that an intercept is included in $\mathbf{x}_i$.

We used LR models to estimate the probability of a positive imaging outcome. The

Table 2.4: Univariable and multivariable LR models predicting the presence of lymph node metastases at diagnosis.

	Univariable logistic regression model		Multivariable logistic regression model		
Variables	**OR (95% CI)**	**p-value**	**OR (95% CI)**	**p-value**	**Overall p-value**
Age at diagnosis (year)	1.02 (0.99 – 1.05)	0.02	1.00 (0.96 – 1.03)	0.83	(0.83)
ln(PSA+1), ng/mL	2.79 (2.21 – 3.54)	< 0.0001	2.16 (1.65 – 2.84)	< 0.0001	(< 0.0001)
Biopsy Gleason score, No. (%)					(0.004)
$\leq 3+4$	Reference		Reference		
$4+3$	15.49 (1.84 – 130.48)	0.01	8.13 (0.91 – 72.94)	0.06	
$8-10$	50.69 (6.96 – 369.16)	< 0.0001	19.72 (2.62 – 148.39)	0.004	
Clinical T stage, No. (%)					(0.0005)
T1	Reference		Reference		
T2	2.05 (1.09 – 3.86)	0.03	8.13 (0.91 – 72.94)	0.06	
T3/4	21.05 (9.52 – 46.56)	< 0.0001	19.72 (2.62 – 148.39)	0.004	
Positive cores, %	35.08 (12.06 – 102.03)	< 0.0001	1.82 (0.47 – 7.01)	0.39	(0.439)

discriminative model for LR is given by:

$$\mathbb{P}(y_i = \pm 1 \,|\, \mathbf{x}_i, \boldsymbol{\beta}) = \frac{1}{1 + e^{-y_i \boldsymbol{\beta}^T \mathbf{x}_i}} \tag{2.1}$$

Under this probabilistic model, the parameter $\boldsymbol{\beta}$ is estimated via maximum likelihood estimation (MLE) by minimizing the conditional negative log-likelihood:

$$-\log \mathbb{L}(\boldsymbol{\beta}) = -\log \prod_{i=1}^{l} \mathbb{P}(y_i = \pm 1 \,|\, \mathbf{x}_i, \boldsymbol{\beta}) = \sum_{i=1}^{l} \log \left(1 + e^{-y_i \boldsymbol{\beta}^T \mathbf{x}_i}\right) \tag{2.2}$$

to obtain well-calibrated predicted probabilities.

2.2.3. Statistical Validation Methods

To evaluate the accuracy of our risk prediction models, we performed both internal and external validation. Internal validation uses the same dataset to develop and validate the model, and external validation uses an independent dataset to validate the model. We used internal validation at early stages of the project when a limited number of samples were available; we subsequently conducted external validation later in the project when a suitable amount of additional data had been collected.

Validating a predictive model using the development sample will introduce bias, known as *optimism*, because the model will typically fit the training dataset better than a new

dataset. Given the intention to implement these guidelines for clinical practice, it was necessary to carefully consider this bias. We used bootstrapping since it is an efficient internal validation technique that addresses this bias to provide more accurate estimates of the performance of a predictive model [52, 71].

Since internal validation has limitations in determining the generalizability of a predictive model [27], we conducted external validation to confirm the validity of the predictive models using new data that was unavailable during the initial model building process. Following is a description of the performance measures that we used to evaluate our models for both forms of validation, as well as a detailed explanation of our two-stage internal and external validation approach.

Performance Metrics

There are two primary aspects in the assessment of the predictive model accuracy: assessment of *discrimination* and *calibration*. Discrimination refers to the ability of the predictive models to distinguish patients with and without metastatic disease, and calibration refers to the agreement between the predicted and observed probabilities.

Discrimination was quantified using the area under the receiver operating characteristics (ROC) curves. The area under the ROC curve (AUC) indicates the likelihood that for two randomly selected patients, one with and one without metastasis, the patient with metastasis has the higher predicted probability of a positive imaging outcome. The AUC provides a single measure of a classifier's performance for evaluating which model is better on average, and assesses the ranking in terms of separation of metastatic patients from cancer-free patients [163]. The larger the AUC the better the performance of the classification model. Figure 2.2 illustrates the ROC curves for the BS and CT scan risk prediction models based on the external validation samples (to be discussed in Section 2.2.3).

We assessed the calibration of the predicted probabilities via the *Brier score*. The Brier score is the average squared difference between the observed label and the estimated probability, calculated as $\sum_{i=1}^{n}(y_i - \mathbb{P}(y_i = 1 \mid \mathbf{x}_i, \boldsymbol{\beta}))/n$, where we assume that n is the size of the sample with which the model is being assessed and $y \in \{0, 1\}$. By definition, the Brier score summarizes both calibration and discrimination at the same time: the square root of the Brier score (root mean squared error) is the expected distance between the observation and the prediction on the probability scale, and lower scores are thus better.

In addition to the Brier score, we evaluated the calibration of the model predictions by

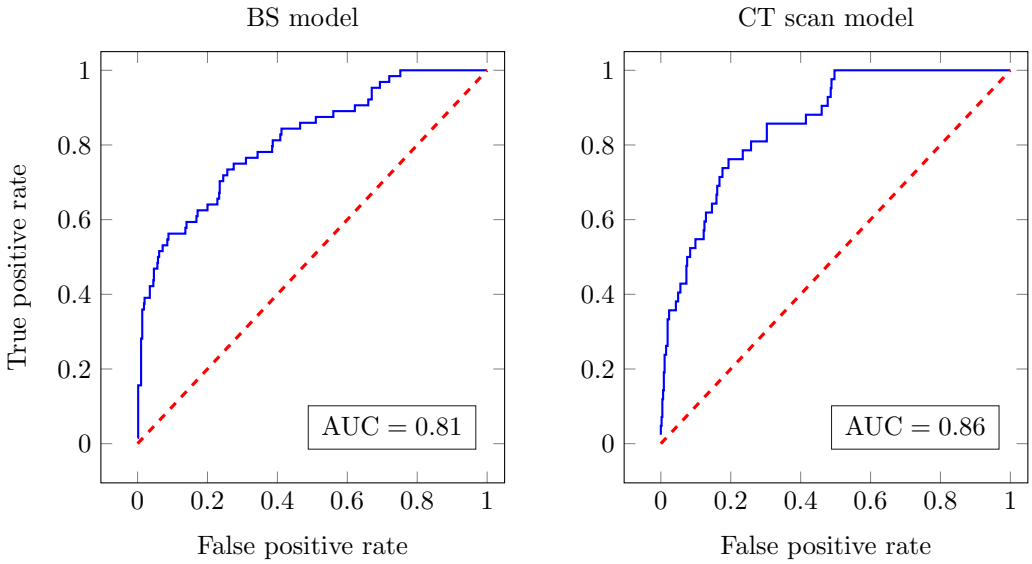

Figure 2.2: The ROC curves for BS and CT scan risk prediction models based on the validation samples.

estimating the slope of the linear predictor of the LR model, known as the *calibration slope* [111]. The linear predictor (LP) is the sum of the regression coefficients multiplied by the patient value of the corresponding predictor (i.e., for patient i, $\text{LP}_i = \mathbf{x}_i\boldsymbol{\beta}$). By definition, the calibration slope is equal to one in the development sample. In an external validation sample, the calibration slope, $\beta_{calibration}$, is estimated using an LR model with the linear predictor as the only explanatory variable (i.e, $\text{logit}(\mathbb{P}(y = 1)) = \alpha + \beta_{calibration}\text{LP}$)[46]. The two estimated parameters in this model, α and $\beta_{calibration}$, are measures of calibration of the LR model in the external validation sample. We can use these parameters to test the hypothesis that the observed proportions in the external dataset are equal to the predicted probabilities from the original model. The slope, $\beta_{calibration}$, is a measure of the direction and spread of the predicted probabilities. Well-calibrated models have a slope of one, indicating predicted risks agree fully with observed frequencies. Models providing overly optimistic predictions will have a slope that is less than one, indicating that predictions of low-risk patients are underestimated and predictions of high-risk patients are overestimated [71, 111].

We assessed the model calibration graphically with calibration plots. We divided the patients into ten, approximately equal-sized groups, according to the deciles of the predicted probability of a positive outcome as derived from the fitted statistical model. Within each

decile, we determined the mean predicted probability (x-axis) and the true fraction of positive cases (y-axis). If the model is well-calibrated, the points will fall near the diagonal line.

Validation Process

In order to determine the internal validity of the predictive models, we used bootstrapping. This involves sampling from the development sample, with replacement, to create a series of random bootstrap samples. In each bootstrap sample, we fit a new LR model and apply this model to the development sample. The expected optimism is then calculated by averaging the differences between the performance of models developed in each of the bootstrap samples (i.e., *bootstrap performance*) and their performance in the development sample (i.e., *test performance*). The optimism is then subtracted from the apparent performance of the original model fit in the development sample to estimate the internally validated performance. Algorithm 1 parallels the approach in [53]. We used this approach to internally validate the model calibration and discrimination.

Following our analysis and guideline development in the initial stages of this project, new validation datasets became available for BS and CT scan, which we used to confirm the validity of the developed predictive models. The inclusion and exclusion criteria, data collection, and clinical variables were identical to those used for the development samples. As part of our external validation, we validated the risk prediction models on these external validation sets using the performance measures described above to estimate discrimination and calibration. We also assessed the external calibration via calibration plots, which we discussed in Section 2.2.3.

2.2.4. Statistical Validation Results

Based on the approach described in Section 2.2.3, we calculated the expected optimism for the AUC, Brier score, and calibration slope (Table 2.5). Comparison of the apparent performance of the risk prediction models with the optimism-corrected performance supported the precision of the model performance estimates in the initial stage of the project.

To assess the generalizability of these models, we evaluated the performance estimates in independent external validation samples collected approximately one year after our initial analysis. Table 2.6 summarizes the results from the external validation of the predictive models. The validation sample for BS included 664 patients, of which 64 (9.64%) had a

Algorithm 1: Bootstrapping Algorithm for Internal Validation.

Input : A predictive model, a development sample of n patients and the number of bootstrap replications m.

Output: The internally validated performance, $P_{validated}$.

1 Estimate the apparent performance of the predictive model, $P_{apparent}$, fit in the development sample.

2 **for** $i = 1, \ldots, m$ **do**

3 Draw a random bootstrap sample of n patients from the development sample with replacement;

4 Fit the logistic regression model to the bootstrap sample and measure the apparent performance in the same sample, $P_{bootstrap}(i)$;

5 Apply the bootstrap model to the development sample and estimate the test performance of this bootstrap model, $P_{test}(i)$;

6 Calculate an estimate of the optimism, $o(i) = P_{bootstrap}(i) - P_{test}(i)$;

7 **end**

8 Estimate the expected optimism:

$$Optimism = \frac{\sum_{i=1}^{m} o(i)}{m}$$

9 Return $P_{validated} = P_{apparent} - Optimism$.

positive outcome with evidence for bone metastasis, and for CT scan scan included 507 patients of which 42 (8.28%) were interpreted as positive for lymph node metastasis. The change in AUC between the internal and external validation for BS and CT scan models was not significant (e.g., 0.01). The increase in the calibration slopes and decrease in the Brier score demonstrate that our models are well-calibrated to the external validation samples. Overall, the expected optimism and optimism-corrected performance as estimated with bootstrapping agreed well with that observed with independent validation samples.

The calibration plots in Figure 2.3 compare observed and predicted probability estimates for the BS and CT scan models. The results show good calibration in the external validation samples. Note that there is only one case in which there is a statistically difference from perfect calibration. The results from internal and external validation demonstrate that the risk prediction models are well-calibrated.

Table 2.5: Bootstrap results for the development samples.

	Development samples	
	BS (n = 416) mean ± $\mathbf{SE_{bootstrap}}$	**CT (n = 643)** mean ± $\mathbf{SE_{bootstrap}}$
Apparent performance		
AUC	0.84	0.89
Brier score	0.075	0.057
Calibration slope	1	1
Bootstrap performance		
AUC	0.86 ± 0.032	0.89 ± 0.021
Brier score	0.073 ± 0.0098	0.056 ± 0.0072
Calibration slope	1	1
Test performance		
AUC	0.83 ± 0.011	0.88 ± 0.0086
Brier score	0.078 ± 0.0016	0.059 ± 0.0014
Calibration slope	0.86 ± 0.18	0.90 ± 0.12
Expected optimism		
AUC	0.023 ± 0.032	0.014 ± 0.022
Brier score	−0.0048 ± 0.0099	−0.0028 ± 0.0072
Calibration slope	0.86 ± 0.18	0.90 ± 0.12
Optimism-corrected performance		
AUC	0.82	0.87
Brier score	0.080	0.060
Calibration slope	0.86	0.90

In the development samples for BS and CT scan, 1000 bootstrap repetitions were used for the calculation of both the mean and standard deviations ($SE_{bootstrap}$).

Table 2.6: Internal and external validation results of the risk prediction models.

	Development samples		Validation samples	
	BS (n = 416)	**CT (n = 643)**	**BS (n = 664)**	**CT (n = 507)**
AUC	0.82	0.87	0.81	0.86
Brier score	0.080	0.060	0.068	0.061
Calibration slope	0.86	0.90	0.99	0.94

Performance measures were found by applying the predictive models fit in the development samples to the validation samples.

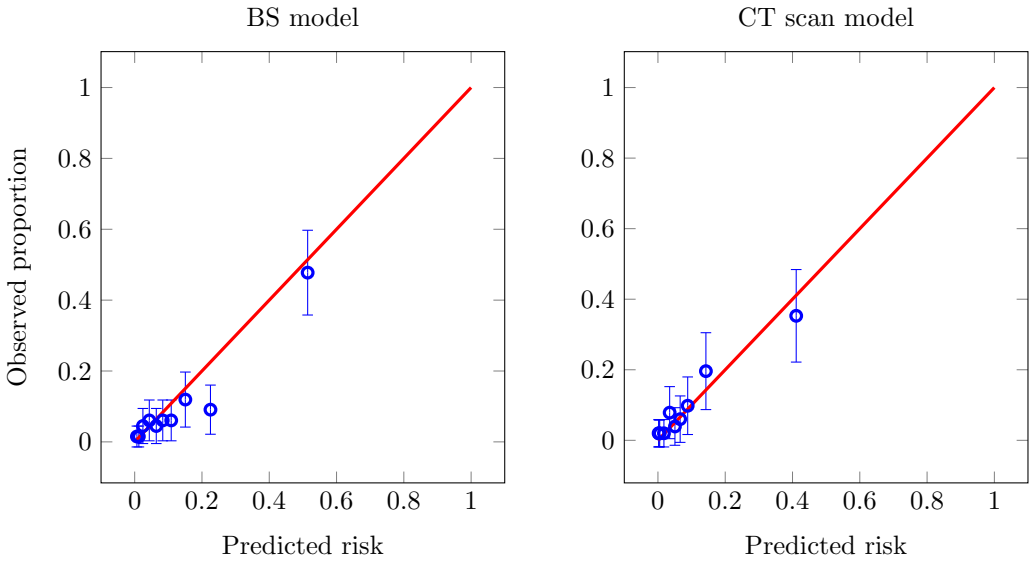

Figure 2.3: Calibration plots for BS and CT scan risk prediction models based on the validation samples.

2.3. Classification Modeling for Metastatic Cancer Detection

This section describes (1) an optimization based approach for the development of classification models that account for missing labels (i.e., imaging outcomes) and class imbalance, and (2) alternative classification modeling techniques that are adapted for advancing the recognition of metastatic patients in imbalanced data.

2.3.1. Background on Classification with Unlabeled and Imbalanced Data

We identify two important challenges regarding the development of classification models in diagnostic medicine: *learning from unlabeled data* and *learning from imbalanced data*. The first challenge, unlabeled data, arises from the fact that in practice not all patients receive a BS or CT scan at diagnosis, which results in a missing data problem. The second challenge, imbalanced data, arises from the fact that a minority of patients has metastatic cancer. To address each of these challenges, we study two machine learning paradigms in this chapter: *semi-supervised* and *cost-sensitive* learning.

Semi-supervised learning aims to improve the learning performance by appropriately

exploiting the unlabeled data in addition to the labeled data [38, 182, 185, 186]. The lack of an assigned clinical class for each patient is the most common situation faced when using observational data in medicine such as in our case. This naturally occurs because patients who appear at high risk of disease receive the gold standard test while patients at lower risk may not.

Class imbalance and cost-sensitive learning are closely related to each other [40, 73, 175]. Cost-sensitive learning aims to make the optimal decision that minimizes the total misclassification cost [51, 57, 103, 106, 162]. Several studies have shown that cost-sensitive methods demonstrated better performance than sampling methods in certain application domains [100, 107, 155, 183].

The use of unlabeled data in cost-sensitive learning has attracted growing attention and many techniques have been developed [67, 95, 98, 104, 130, 131]. To our knowledge, however, there has not been an attempt to apply both semi-supervised and cost-sensitive learning to improve cancer diagnosis (see the literature reviews in [48] and [90]). In this chapter, we focus on using kernel logistic regression (KLR) to address unequal costs and utilize unlabeled data simultaneously based on a novel extension of the framework for data-dependent geometric regularization [15].

2.3.2. Classification Models

We begin by introducing our approach for the construction of a classification model that exploits data of patients with missing imaging outcomes and improves the identification performance on the minority class by incorporating unequal costs in the classification loss.

Regularization is a key method for obtaining smooth decision functions and thus avoiding *over-fitting* to the training data, which is widely used in machine learning [15, 58]. In this context, we represent a classifier as a mapping $\mathbf{x} \mapsto \text{sign}\,(f(\mathbf{x}))$, where f is a real-valued function $f : \mathbb{R}^d \to \mathbb{R}$, sometimes called a *decision function*. We adopt the convention $\text{sign}(0) = -1$. A general class of regularization problems estimates the unknown function f by minimizing the functional:

$$\min_{f \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} L(y_i, f(\mathbf{x}_i)) + \gamma_{\mathcal{H}} \|f\|^2_{\mathcal{H}} \tag{2.3}$$

where $L(y, f(\mathbf{x}))$ is the loss function, $\|\cdot\|_{\mathcal{H}}$ is the Euclidean norm in a high-dimensional (possibly infinite-dimensional) space of functions $\mathcal{H}$. The space $\mathcal{H}$ is defined in terms of a

positive definite *kernel* function $\mathbf{K} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Conditions for a function to be a kernel are expressed by Mercer Theorem; in particular, it must be expressed as an inner product and must be positive semidefinite [147]. The parameter $\gamma_{\mathcal{H}} \geq 0$ is called the regularization parameter and is a fixed, user-specified constant controlling the smoothness of f in $\mathcal{H}$. By the Representer Theorem [88], the minimizer $f^*(\mathbf{x})$ of (2.3) has the form:

$$f^*(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i^* \mathbf{K}(\mathbf{x}, \mathbf{x}_i) \tag{2.4}$$

As a consequence, (2.3) is reduced from a high-dimensional optimization problem in $\mathcal{H}$ to an optimization problem in $\mathbb{R}^l$; where the decision variable is the coefficient vector $\boldsymbol{\alpha}$. The same algorithmic framework is utilized in many regression and classification schemes such as support vector machines (SVM) and regularized least squares [15].

The purpose of optimizing in the higher-dimensional space $\mathcal{H}$ is to consider decision functions that are linear in $\mathcal{H}$, but which may represent nonlinear relationships in the feature space $\mathbb{R}^d$. The kernel also implicitly defines a function $\Phi : \mathbb{R}^d \to \mathcal{H}$ that maps a data point $\mathbf{x}$ in the original feature space $\mathbb{R}^d$ to a vector $\Phi(\mathbf{x})$ in the higher dimensional feature space $\mathcal{H}$. Although explicit knowledge of the transformation $\Phi(\cdot)$ is not available, dot products in $\mathcal{H}$ can be substituted with the kernel function through the *kernel trick*, that is, $\langle \Phi(\mathbf{x}) , \Phi(\mathbf{x}') \rangle = \mathbf{K}(\mathbf{x}, \mathbf{x}')$.

Scaling of (2.2) by a factor of $1/n$ establishes the equivalence between LR estimated by maximum likelihood and empirical risk minimization with *logistic loss*, given as $L(y, f(\mathbf{x})) = \ln(1 + \exp^{-yf(\mathbf{x})})$, in (2.3), where $f(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$ and $\boldsymbol{\beta} \in \mathbb{R}^d$ is a d-dimensional vector of patient attributes. This can be seen as the special case $\mathbf{K}(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle$, corresponding to $\mathcal{H} = \mathbb{R}^d$ and an identity mapping $\Phi(\mathbf{x}) = \mathbf{x}$. However, LR linearity may be an obstacle to handling highly nonlinearly separable data sets. In such cases, nonlinear classification models can achieve superior discrimination accuracy compared to linear models. To include nonlinear decision boundaries in our problem, we extend the construction from LR to KLR by incorporating a non-linear feature mapping into the decision function: $f(\mathbf{x}) = \Phi(\mathbf{x})\boldsymbol{\beta}$ [102, 184]. The optimization problem becomes as follows:

$$\min_{\boldsymbol{\beta} \in \mathcal{H}} \sum_{i=1}^{l} \log(1 + \exp(-y_i \langle \boldsymbol{\beta}, \Phi(\mathbf{x}_j) \rangle)) + \frac{\lambda}{2} \|\boldsymbol{\beta}\|^2, \tag{2.5}$$

where $\boldsymbol{\beta} \in \mathcal{H}$ is the parameter we want to estimate. By (2.4) and the kernel trick, the

minimizer of (2.5) admits a representation of the form $\boldsymbol{\beta} = \sum_{i=1}^{l} \alpha_i \Phi(\mathbf{x}_i)$. Thus, we can write (2.5) as:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^l} \sum_{i=1}^{l} \log(1 + \exp(-y_i(\mathbf{K}\boldsymbol{\alpha})_i)) + \frac{\lambda}{2} \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} \tag{2.6}$$

where $\mathbf{K}$ is the kernel matrix of imaged patients given as $\mathbf{K} = (\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^{l}$ with $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$ and $(\mathbf{K}\boldsymbol{\alpha})_i$ stands for the i-th element of the vector $\mathbf{K}\boldsymbol{\alpha}$.

In order to address the issue of missing data for patients who did not receive a BS or CT scan, we use the Laplacian semi-supervised framework proposed by [15], which extends the classical framework of regularization given in (2.3) by incorporating unlabeled data via a regularization term in addition to the $\mathcal{H}$ norm. Assume a given set of l imaged patients $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ and a set of u unimaged patients $\{\mathbf{x}_j\}_{j=l+1}^{j=l+u}$. In the sequel, let us redefine $\mathbf{K}$ as an $(l+u) \times (l+u)$ kernel matrix over imaged and unimaged patients given by $\mathbf{K} = (\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^{l+u}$ with $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$. Since we do not know the marginal distribution which unimaged patients are drawn from, the empirical estimates of the underlying structures (i.e., clusters) inherent in unimaged data is encoded as a graph whose vertices are the imaged and unimaged patients and whose edge weights represent appropriate pairwise similarity relationships between patients [148].

The concept underlying this new regularization comes from *spectral clustering*, which is one of the most popular clustering algorithms [171]. To define a graph Laplacian, we let G be a weighted graph with vertices corresponding to all patients. When the data point $\mathbf{x}_i$ is among the k-nearest neighbors of $\mathbf{x}_j$, or $\mathbf{x}_j$ is among those of $\mathbf{x}_i$, these two vertices are connected by an edge, and a nonnegative weight w_{ij} representing the similarity between the points $\mathbf{x}_i$ and $\mathbf{x}_j$ is assigned. The weighted *adjacency matrix* of graph G is the symmetric $(l+u) \times (l+u)$ matrix $\mathbf{W}$ with the elements $\{w_{ij}\}_{i,j=1}^{l+u}$, and the *degree matrix* $\mathbf{D}$ is the diagonal matrix with the degrees $d_1, \ldots, d_{l+u}$ on the diagonal, given as $d_i = \sum_{j=1}^{l+u} w_{ij}$. Defining $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_{l+u})]^T$, and $\mathbf{L}$ as the Laplacian matrix of the graph given by $\mathbf{L} = \mathbf{D} - \mathbf{W}$, we consider the following optimization problem:

$$f^* = \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} L\left(y_i, f(\mathbf{x}_i)\right) + \gamma_{\mathcal{H}} \|f\|_{\mathcal{H}}^2 + \gamma_{\mathcal{M}} \mathbf{f}^{\mathbf{T}} \mathbf{L} \mathbf{f} \tag{2.7}$$

where $\gamma_{\mathcal{H}}$ and $\gamma_{\mathcal{M}}$ are the regularization parameters that control the $\mathcal{H}$ norm and the *intrinsic norm*, respectively. In this context, the Laplacian term forces to choose a deci-

sion function f that produces similar outputs for two patients with high similarity, i.e., connected by an edge with a high weight, regardless of their imaging status.

For the purposes of this chapter, we will consider asymmetric loss functions with unequal misclassification costs so that the cost of misclassifying a patient with metastasis outweighs the cost of misclassifying a cancer-free patient. We can formulate the cost-sensitive classification loss given by $L_\delta : \{-1, 1\} \times \mathbb{R} \to [0, \infty]$ with cost parameter $\delta \in (0, 1)$ as:

$$L_\delta = \delta \mathbb{1}_{\{y=1\}} L_1(f(\mathbf{x})) + (1-\delta)\mathbb{1}_{\{y=-1\}} L_{-1}(f(\mathbf{x})) \tag{2.8}$$

where we refer to L_1 and L_{-1} as the *partial losses* of L ([144]). In KLR, the partial losses can be defined as $L_1(f(\mathbf{x})) = \log(1 + e^{-f(\mathbf{x})})$ and $L_{-1}(f(\mathbf{x})) = \log(1 + e^{f(\mathbf{x})})$. From (2.8), the cost-sensitive optimization problem can then be formulated as:

$$\begin{aligned} f^* &= \underset{f \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{l} \sum_{i=1}^{l} \left[\delta \mathbb{1}_{\{y_i=1\}} \log\left(1 + e^{-f(\mathbf{x}_i)}\right) + (1-\delta)\mathbb{1}_{\{y_i=-1\}} \log\left(1 + e^{f(\mathbf{x}_i)}\right) \right] \\ &\quad + \gamma_{\mathcal{H}} \|f\|^2_{\mathcal{H}} + \gamma_{\mathcal{M}} \mathbf{f^T L f} \end{aligned} \tag{2.9}$$

We refer to the optimization problem in (2.9) as cost-sensitive Laplacian kernel logistic regression (Cos-LapKLR). The extensions of standard regularization algorithms by solving the optimization problems (posed in (2.3)) for different choices of cost function L and regularization parameters $\gamma_{\mathcal{H}}$ and $\gamma_{\mathcal{M}}$ have been developed [15]. We extend their work by formulating the logistic loss for KLR in terms of partial losses to adjust for class imbalance while exploiting the information from unimaged patients.

As before, the Representer Theorem can be used to show that the solution to (2.9) has an expansion of kernel functions over both the imaged and unimaged given as $f^*(\mathbf{x}) = \sum_{i=1}^{l+u} \alpha_i^* \mathbf{K}(\mathbf{x}_i, \mathbf{x})$. Let $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_L^T, \boldsymbol{\alpha}_U^T]^T$ be the $l+u$-dimensional variable with $\boldsymbol{\alpha}_L = [\alpha_1, \ldots, \alpha_l]^T$ and $\boldsymbol{\alpha}_U = [\alpha_{l+1}, \ldots, \alpha_{l+u}]^T$, and $\mathbf{K}_L \in \mathbb{R}^{l \times l}$ be the kernel matrix for imaged patients. In order to express (2.9) in terms of the variable $\boldsymbol{\alpha}$, we define $\mathbf{P}_L = [\mathbf{I}_{l\times l} \quad \mathbf{0}_{l\times u}]$ and substitute $\boldsymbol{\alpha}_L$ as $\boldsymbol{\alpha}_L = \mathbf{P}_L \boldsymbol{\alpha}$. Let $H(\boldsymbol{\alpha})$ denote the objective function with respect to $\boldsymbol{\alpha}$. Introducing linear mappings, (2.9) can then be equivalently re-written in a finite dimensional form as:

$$\begin{aligned} H(\boldsymbol{\alpha}) &= \min_{\boldsymbol{\alpha} \in \mathbb{R}^{l+u}} \frac{1}{2l} \Big[\delta \mathbf{1} \left(\mathbf{1} + \mathbf{y}\right)^T \log\left(\mathbf{1} + e^{-(\mathbf{K}_L \mathbf{P}_L \boldsymbol{\alpha})}\right) + \\ &\quad + (1-\delta)\mathbf{1}\left(\mathbf{1} - \mathbf{y}\right)^T \log\left(\mathbf{1} + e^{(\mathbf{K}_L \mathbf{P}_L \boldsymbol{\alpha})}\right) \Big] + \gamma_{\mathcal{H}} \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} + \gamma_{\mathcal{M}} \boldsymbol{\alpha}^T \mathbf{K L K} \boldsymbol{\alpha} \end{aligned} \tag{2.10}$$

The outline of the algorithm we propose for solving Cos-LapKLR is given in Algorithm 2. It is natural to use the Newton-Raphson method to fit the Cos-LapKLR since (2.10) is strictly convex. However, the drawback of the Newton-Raphson method is that in each iteration an $(u+l)\times(u+l)$ matrix needs to be inverted. Therefore, the computational cost is $O((u+l)^3)$. When $(u+l)$ becomes large, this can become prohibitively expensive. In order to reduce the cost of each iteration of the Newton-Raphson method, we implemented one of the most popular *quasi-Newton* methods, the so-called Broyden-Fletcher-Goldfarb-Shanno (BFGS) method. It approximates the Hessian instead of explicitly calculating it at each iteration [49]. We used the limited-memory BFGS (LM-BFGS), which is an extension to the BFGS algorithm which uses a limited amount of computer memory [35].

Algorithm 2: Cost-sensitive Laplacian Kernel Logistic Regression (Cos-LapKLR).

Input : l labeled examples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$, u unlabeled examples $\{\mathbf{x}_j\}_{j=l+1}^{l+u}$.
Output: Estimated function $f : \mathbb{R}^{(l+u)} \to \mathbb{R}$.
Step 1: Construct the data adjacency graph with $(l+u)$ nodes and compute the edge weights w_{ij} by k nearest neighbors.
Step 2: Choose a kernel function and compute the kernel matrix $\mathbf{K} \in \mathbb{R}^{(l+u)\times(l+u)}$.
Step 3: Compute the graph Laplacian matrix: $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D} = \text{diag}\,(d_1, \ldots, d_{l+u})$ and $d_i = \sum_{j=1}^{l+u} w_{ij}$.
Step 4: Choose the regularization parameters $\gamma_{\mathcal{H}}$, $\gamma_{\mathcal{M}}$, and the cost parameter δ.
Step 5: Compute $\boldsymbol{\alpha}^*$ using (2.10) together with the LM-BFGS algorithm.
Step 6: Output function $f^*(\mathbf{x}) = \sum_{i=1}^{l+u} \alpha_i^* \mathbf{K}(\mathbf{x}_i, \mathbf{x})$.

In addition to Cos-LapKLR, we implemented and tested several other well-known classification models including LR, Random forests (RF) [29], SVM [169], and AdaBoost [62]. As discussed earlier in this section, LR can be estimated by minimizing the logistic loss. Hence, we adopted asymmetric loss functions in LR, which we refer to as cost-sensitive logistic regression (Cos-LR), in a similar manner as proposed for KLR to counter the effect of class imbalance due to having fewer patients with metastasis. Since the logistic loss minimization problem in Cos-LR is convex, LM-BFGS was applied to this problem as well.

Similar to Cos-LapKLR and Cos-LR, the SVM hinge loss can be extended to the cost-sensitive setting by introducing penalties for misclassification [170]. The regularization parameter C in cost-sensitive support vector machines (Cos-SVM) corresponds to the misclassification cost which involves two parts, i.e., the cost of misclassifying negative class into positive class and the cost of misclassifying positive class into negative class. In this

work, the cost of misclassifying negative class as positive is set to C, whereas the cost of misclassifying positive class into negative class is set to $C \times \delta/(1-\delta)$, where $\delta \in (0,1)$.

To remedy the class imbalance problem with RF and AdaBoost, different data sampling techniques were employed in the experimental evaluation, such as random oversampling (ROS), random undersampling (RUS), and the combination of both methods. ROS and RUS are non-heuristic methods that are initially included in this evaluation as baseline methods. The drawback of resampling is that undersampling can potentially lose some useful information, and oversampling can lead to overfitting [39]. To overcome these limitations, we also implemented advanced balancing methods for comparison. A brief discussion of the concepts underlying these methods is provided in Appendix A.1.

Classification Model Results

We adopted 2-fold cross validation (CV) in the model training process. The radial basis function kernel of the form $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$ was used, where γ is the kernel parameter. The continuous attributes were normalized to a mean of zero and standard deviation of one. All models were built and evaluated with Python 2.7.11 on a HP Z230 work station with an Intel Xeon E31245W (3.4GHz) processor, 4 cores, and 16 GB of RAM. We used the `scipy.optimize` package in Python as the optimization solver.

Our goal was to obtain a higher identification rate for metastatic patients without greatly compromising the classification of patients without metastasis. Therefore, we created trade-off curves to determine Pareto optimal models based on sensitivity and specificity. Sensitivity, or true positive rate, indicates the accuracy on the positive class; specificity, or true negative rate, indicates the accuracy on the negative class. In the concept of Pareto optimality, a model is considered *dominated* if there is another model that has a higher sensitivity and a higher specificity. For cost-sensitive classification models, we created Pareto frontier graphs consisting of the non-dominated models for varying choices of cost parameter based on 2-fold CV performance. We conducted experiments for $\delta \in \{0, 1\}$; however, we report results for $\delta \in \{0.90, 0.91, \ldots, 0.99\}$ to be consistent with the goals of the project and the perspective of stakeholders who weigh the misclassification of patients with cancer much higher than patients without cancer.

Following the approach of [80] recommended for SVM, the values of the remaining hyperparameters for Cos-LapKLR, Cos-LR and Cos-SVM models were chosen from a range of different values after 2-fold CV at different cost setups. For Cos-LapKLR, candidate values

for the regularization parameters $\gamma_{\mathcal{H}}$ and $\gamma_{\mathcal{M}}$ are chosen from the set $\{2^i \mid -13, -11, \ldots, 3\}$, the kernel parameter γ from $\{2^i \mid -9, -7, \ldots, 3\}$, and the nearest neighbor parameter k from $\{3, 5\}$. For Cos-LR, candidate values for the regularization parameter λ is chosen from the set $\{2^i \mid -13, -11, \ldots, 3\}$. For Cos-SVM, candidate values for the regularization parameter C is chosen from the set $\{2^i \mid -5, -3, \ldots, 15\}$ and the kernel parameter γ from $\{2^i \mid -15, -13, \ldots, 3\}$. We defined the weight matrix $\mathbf{W}$ by k-nearest neighbor for Cos-LapKLR models as follows [15]:

$$w_{ij} = \begin{cases} e^{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2}, & \text{if } \mathbf{x}_i, \mathbf{x}_j \text{ are neighbors} \\ 0, & \text{otherwise} \end{cases}$$

We applied the Pareto frontier based approach to select the optimal classifiers for each of these methods for distinguishing patients with metastasis at different cost setups during the training process.

For RF, we used the nominal values recommended by [61] for the number of trees to grow (500) and minimum node size (5). For AdaBoost, we used single-split trees with two nodes as the base learner, since this was shown to yield good performance of AdaBoost [62, 141]. We performed 10 independent runs of 2-fold CV to eliminate bias that could occur as a result of the random partitioning process. For conciseness, the detailed results from these experiments are presented in Appendix A.1. In the remainder of this section, we summarize results for the cost-sensitive methods (i.e., Cos-LapKLR, Cos-LR and Cos-SVM).

Our initial experiments explored how the cost ratio, δ, affects the classification performance of the cost-sensitive methods as the cost ratio is changing. To illustrate the effect of asymmetrical logistic loss functions, we present Pareto frontier graphs based on sensitivity and specificity for the symmetric ($\delta = 0.5$) and asymmetric ($\delta = 0.95$) cases. Figure 2.4 shows that increasing δ can improve sensitivity significantly without greatly sacrificing specificity. We observed the same trend for Cos-LapKLR models predicting CT scan outcomes, and for Cos-LR and Cos-SVM models for both BS and CT scan with respect to increasing values of δ.

Our next set of experiments, in Figure 2.5, illustrates the impact of increasing the penalty of L_1 loss on the discriminative ability of the LR and Lap-KLR models for predicting BS outcomes. For simplicity, we present the results for only two dimensions (ln(PSA + 1) and age). We see that higher penalty on L_1 loss increases the region of $\mathbb{P}(y = 1 \mid \mathbf{x})$, corresponding to patients with predicted outcome $\hat{y} = 1$, i.e., $f(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta} \geq 0$, and thus,

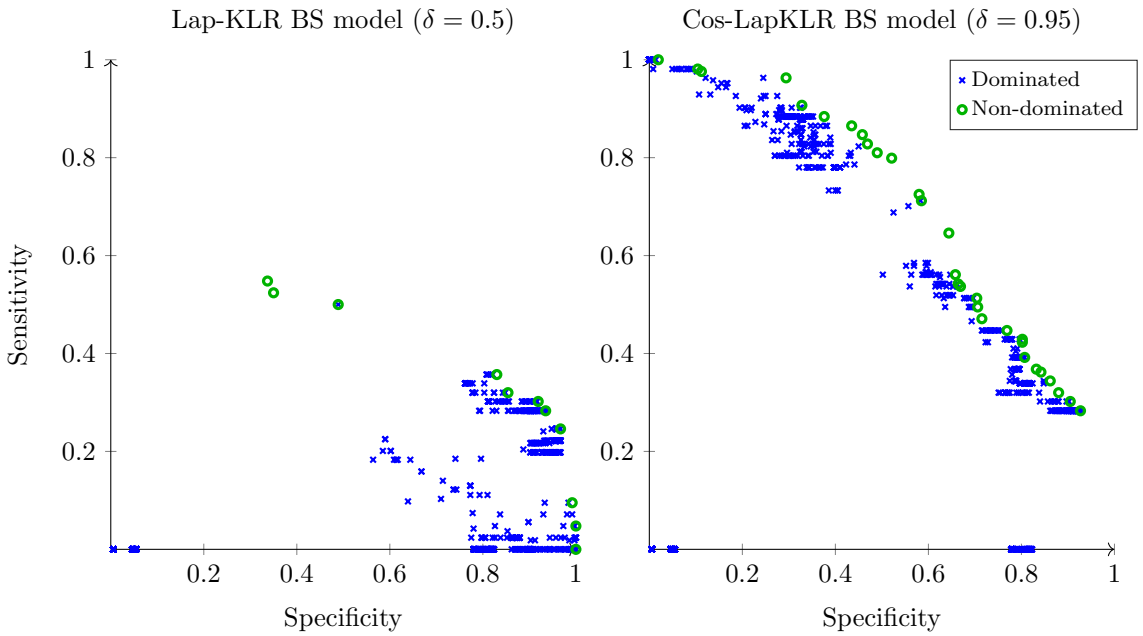

Figure 2.4: Pareto frontier graphs demonstrating the efficient frontiers based on sensitivity and specificity for Laplacian models predicting BS outcomes.

sensitivity of the classification rule increases while specificity decreases with increasing values of δ.

2.4. Bias-corrected Performance of Imaging Guidelines

The results presented in Section 6 for the sensitivity and specificity of alternative classification models are systemically biased since they are based on only the patients who received BS or CT scan at diagnosis. This section provides some background on this problem of verification bias and presents results for the application of the proposed methodology we used to correct for this bias.

2.4.1. Background

Standard inferential procedures rely on several assumptions concerning study design such as the existence of a reference test, usually referred to as a *gold standard*, a procedure that is known to be capable of classifying an individual as diseased or nondiseased. In practice, gold standard tests are often invasive and may be expensive (e.g., BS or CT scan scan are

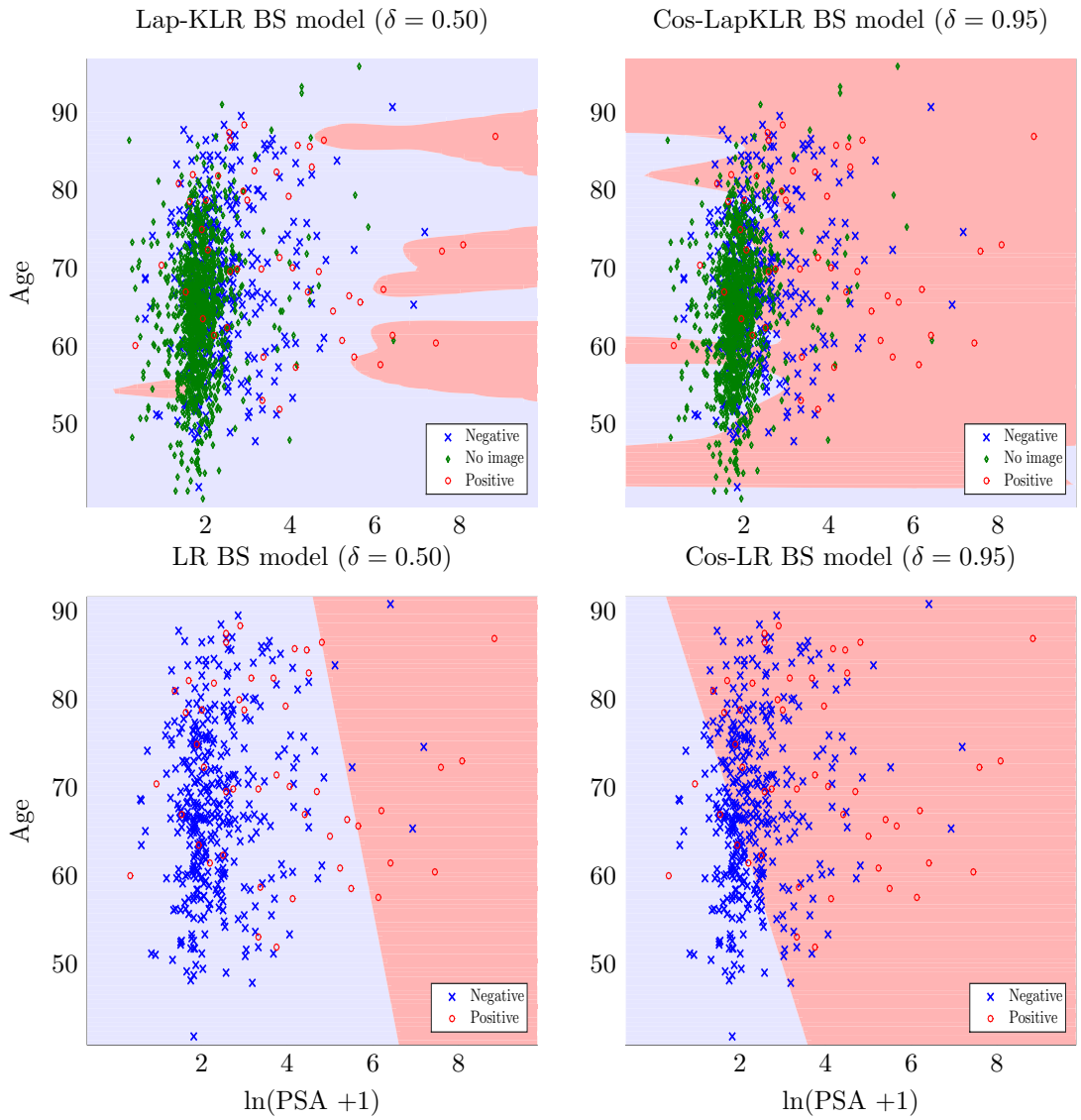

Figure 2.5: The impact of unequal misclassification costs on the decision boundaries of Cos-LR and Cos-LapKLR.

gold standard tests for detecting metastatic cancer). As a result, the true disease status is generally not known for some patients in a study cohort. Moreover, the decision to verify presence of the disease with a gold standard test is often influenced by individual patient risk factors. Patients who appear to be at high risk of disease may very likely to be offered the gold standard test, whereas patients who appear to be at lower risk are less likely. Thus, if only patients with verified disease status are used to assess the diagnostic accuracy of the test, the resulting model is likely to be biased. This bias is referred to as

verification bias (or *work-up bias*) [13]. This can markedly increase the apparent sensitivity of the test and reduce its apparent specificity [13, 89, 121].

Several approaches have been proposed to address the problem of verification bias [180, 181]. The correction methods proposed recently have been mainly focused on treating the verification bias problem as a missing data problem, in which the true disease status is missing for patients who were not selected for the gold standard verification. In the proposed missing data techniques, inferences depend on the nature of incompleteness. In the usual terminology, data are missing at random (MAR) when the mechanism resulting in its omission depends only on the observed data [97]. Thus, given the test results and patient covariates, the missingness mechanism does not depend on the unobserved data (i.e., metastatic disease status). Data are said to be missing completely at random if the missing data mechanism doesn't depend on the observed or missing data.

To obtain unbiased estimates of sensitivity and specificity, Begg and Greenes (B&G) developed a method based on MLE [14]. This method uses the observed proportion of patients with and without the disease among the verified patients to calculate the expected proportion among nonverified patients. The two are then combined to obtain a complete two-by-two table, as if all patients had received the gold standard test. We used this method to correct for verification bias in the assessment of imaging guidelines. The underlying assumption in this method is that the available covariates were the only factors that influenced selection of patients recommended for imaging (i.e., MAR assumption). This is a reasonable assumption given that the MUSIC data repository includes all standard covariates related to metastatic PCa risk.

In this framework, we define the "test" to be the outcome of applying a given guideline (G), where "+" and "−", denote whether a patient is recommended to receive an imaging test or not under the guideline G, respectively. The uncorrected sensitivity and specificity are defined as:

$$\text{Sensitivity} = \mathbb{P}(\text{G}+ \mid \text{Disease present}), \quad \text{Specificity} = \mathbb{P}(\text{G}- \mid \text{Disease not present})$$

Using *Bayes*'s rule, we estimate the sensitivity and specificity of the guideline as follows:

$$\text{Sensitivity} = \mathbb{P}(\text{G+} \mid \text{Disease present}) = \frac{\mathbb{P}(\text{Disease present} \mid \text{G+})\mathbb{P}(\text{G+})}{\mathbb{P}(\text{Disease present})}$$

$$\text{Specificity} = \mathbb{P}(\text{G−} \mid \text{Disease not present}) = \frac{\mathbb{P}(\text{Disease not present} \mid \text{G−})\mathbb{P}(\text{G−})}{\mathbb{P}(\text{Disease not present})}$$

where $\mathbb{P}(\text{Disease present})$ and $\mathbb{P}(\text{Disease not present})$ can be calculated as follows:

$$\mathbb{P}(\text{Disease present}) = \mathbb{P}(\text{Disease present}|\text{G+})\mathbb{P}(\text{G+}) + \mathbb{P}(\text{Disease present}|\text{G−})\mathbb{P}(\text{G−})$$

$$\mathbb{P}(\text{Disease not present}) = \mathbb{P}(\text{Disease not present}|\text{G+})\mathbb{P}(\text{G+}) + \mathbb{P}(\text{Disease not present}|\text{G−})\mathbb{P}(\text{G−})$$

Thus, to estimate the sensitivity and specificity of each guideline, we need to calculate $\mathbb{P}(\text{Disease present} \mid \text{G+})$, $\mathbb{P}(\text{Disease not present} \mid \text{G−})$, $\mathbb{P}(\text{G+})$, and $\mathbb{P}(\text{G−})$. To estimate $\mathbb{P}(\text{Disease present} \mid \text{G+})$ and $\mathbb{P}(\text{Disease not present} \mid \text{G−})$, we first separate the entire population (with and without imaging results) into two categories: (1) those patients with G+ and (2) those patients with G−. To calculate $\mathbb{P}(\text{Disease present} \mid \text{G+})$, we apply the risk prediction model from Section 2.2 to estimate the mean probability that the disease is present in the G+ category of patients. To calculate $\mathbb{P}(\text{Disease not present} \mid \text{G−})$, we apply the risk prediction model to estimate the mean probability that the disease is not present in the G− category of patients. We further obtain unbiased estimates of $\mathbb{P}(\text{G+})$ and $\mathbb{P}(\text{G−})$ as the proportion of the population in G+ and G−. We then use these estimates to calculate the sensitivity and specificity using the formula defined above.

2.4.2. Bias-Corrected Results

There are several published clinical guidelines for BS and CT scans based on patient PCa characteristics. These guidelines are summarized in Table 2.7. Table 2.8 presents the bias-corrected results for these published guidelines. We found that the estimates of uncorrected sensitivity are significantly higher than the bias-corrected estimates, while uncorrected values for specificity underestimate the true specificity of the existing guidelines. For example, the uncorrected sensitivity and specificity of the American Urology Association (AUA) guideline [161] for recommending BS were 97.92% and 43.48%, respectively, whereas the bias-corrected values were 81.18% and 82.05%, respectively, on the development samples.

Table 2.7: Published clinical guidelines for recommending BS and CT scan.

Bone scan		CT	
Clinical guidelines	**Recommend imaging if any of these:**	**Clinical guidelines**	**Recommend imaging if any of these:**
EAU [115]	GS $\geq$ 8 cT3/T4 disease PSA $>$ 10 ng/ml Symptomatic	**EAU** [74]	GS $\geq$ 8 cT3/T4 disease PSA $>$ 10 ng/ml Symptomatic
AUA [161]	GS $\geq$ 8 PSA $>$ 20 ng/ml Symptomatic	**AUA** [36]	GS $\geq$ 8 PSA $>$ 20 ng/ml cT3/T4 disease Symptomatic
NCCN [117]	cT1 disease & PSA $>$ 20 ng/ml cT2 disease & PSA $>$ 10 ng/ml GS $\geq$ 8 cT3/T4 disease Symptomatic		
Briganti's CART [30]	GS $\geq$ 8 $\geq$ cT2 disease & PSA $>$ 10 ng/ml Symptomatic		

EAU: European Urological Association; AUA: American Urological Association; NCCN: National Comprehensive Cancer Network; CART: classification and regression tree.

Table 2.8: Performance characteristics of the published guidelines before and after correcting for verification bias.

	Development samples				Validation samples			
	Uncorrected		Bias-corrected		Uncorrected		Bias-corrected	
Clinical guidelines	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**
Bone scan								
EAU [115]	97.92	33.97	84.45	75.66	98.44	21.00	89.13	65.98
AUA [161]	97.92	43.48	81.18	82.05	96.88	36.00	85.82	74.84
NCCN [117]	97.92	40.76	82.23	80.86	96.88	32.67	86.94	73.23
Briganti's CART [30]	89.58	45.38	79.31	83.28	93.75	37.67	85.07	75.99
CT scan								
EAU [74]	98.39	36.49	89.92	74.43	100.00	32.04	87.47	75.47
AUA [36]	96.77	49.23	87.21	82.53	100.00	45.81	83.91	83.49

The numbers are the percentages.

We applied the bias-correction method on the optimized classification models of Section 2.3. Figure 2.6 shows the Pareto frontier graph consisting of all the imaging guidelines.

The results indicate that the classification rules obtained using the methods of Section 2.3 can provide a diverse range of classification rules that vary on the basis of sensitivity and specificity. All of the published guidelines have high sensitivity for BS; however they vary more significantly in specificity. For CT scan, the AUA guideline had higher sensitivity and moderately lower specificity. For BS, all of the published guidelines were at the Pareto frontier. For CT scan, all of the published guidelines were dominated by classification rules described in Section 2.3 but were all close to the Pareto frontier.

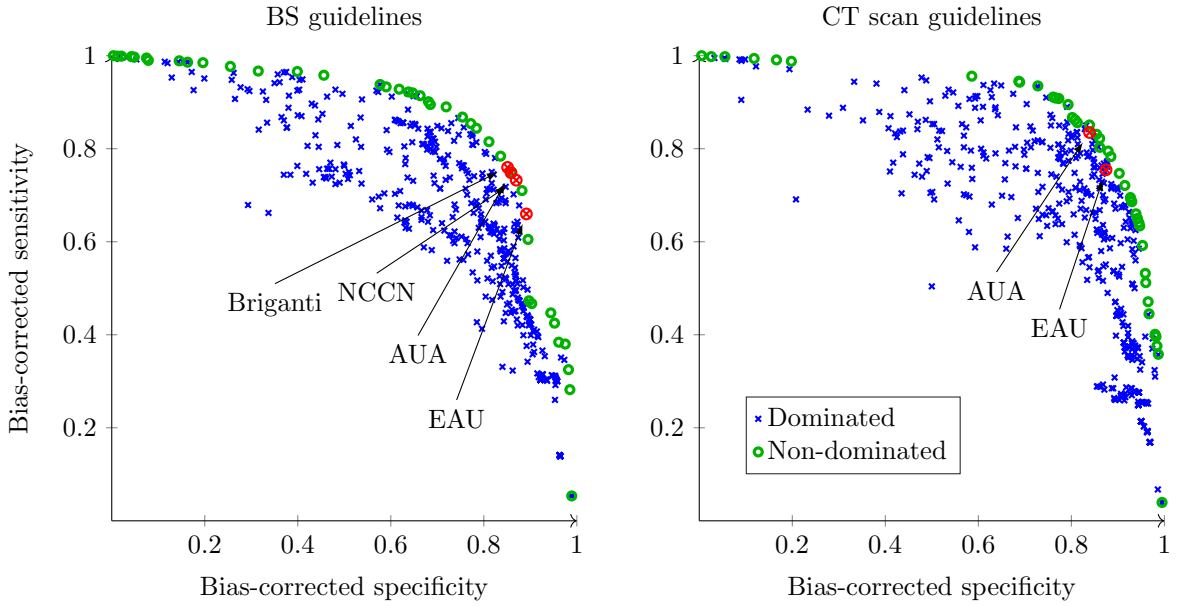

Figure 2.6: Pareto frontier graphs demonstrating the efficient frontiers for the bias-corrected accuracy of the imaging guidelines for BS and CT scan estimated on the validation samples.

To further assess the performance of the statistical methods, we determine the proportions of the non-dominated models for each method based on these two competing criteria. Table 2.9 shows that there is no single classification modeling technique that is sufficient with respect to the estimated number of positive imaging tests missed and the number of negative imaging tests. Thus, underscoring the importance of employing multiple methods for optimization of classification rules.

Table 2.9: Proportions of classification modeling techniques that are non-dominated with respect to the bias-corrected accuracy.

Statistical models	Bone scan (n = 40)	CT (n = 42)
Cos-LapKLR	7.50	30.95
Cos-LR	47.50	0.00
Cos-SVM	27.50	40.48
RF	17.50	9.52
AdaBoost	0.00	19.05

The numbers are the percentages.

2.4.3. Patient Centered Criteria

In working with the MUSIC collaborative we found that interpreting the results was easier when they were presented in terms of more patient-centered health outcomes. Therefore, we considered two important criteria: expected number of positive outcomes missed and expected number of negative studies. These estimates around the impact of specific guideline implementation can provide useful information for clinicians, specialty societies, and other stakeholders seeking a satisfactory tradeoff between the benefits and harms of using these imaging tests for the staging of patients with newly-diagnosed PCa.

To define the criteria to be considered in the objective function, let $p_i = \mathbb{P}(y_i = 1 \mid \mathbf{x}_i, \boldsymbol{\beta})$ be the probability that patient i with attributes $\mathbf{x}_i$ would have a positive imaging outcome, where $i = 1, \ldots, n$, and is estimated from an LR model. Let g_i be an indicator variable defined as:

$$g_i = \begin{cases} 1, & \text{if the guideline is satisfied} \\ 0, & \text{otherwise} \end{cases}$$

If Z^+ denotes a random variable for the number of positive outcomes missed and Z^- a random variable for the number of negative outcomes, then the criteria can expressed as:

$$\mathbb{E}[Z^+] = \sum_{i=1}^{n} p_i \left(1 - g_i\right), \quad \mathbb{E}[Z^-] = \sum_{i=1}^{n} \left(1 - p_i\right) g_i$$

where $\mathbb{E}$ is the expectation operator. Assuming the goal is to find an optimal guideline that minimizes an *unweighted* function of these two competing criteria, the optimization

model can be expressed as:

$$\begin{aligned} \min \quad & Z(g) = [Z^+(g),\ Z^-(g)] \\ \text{subject to} \quad & g \in G \end{aligned}$$

where G is the set of all imaging guidelines consisting of the published clinical guidelines and the non-dominated classification rules from Section 2.4.2. For each $g \in G$, we calculated the expected number of positive imaging outcomes missed and the expected number of negative imaging outcomes based on the validation samples. Figure 2.7 shows that the published guidelines are very close to the efficient frontier for both BS and CT scan, while also achieving a missed metastasis rate $< 1\%$.

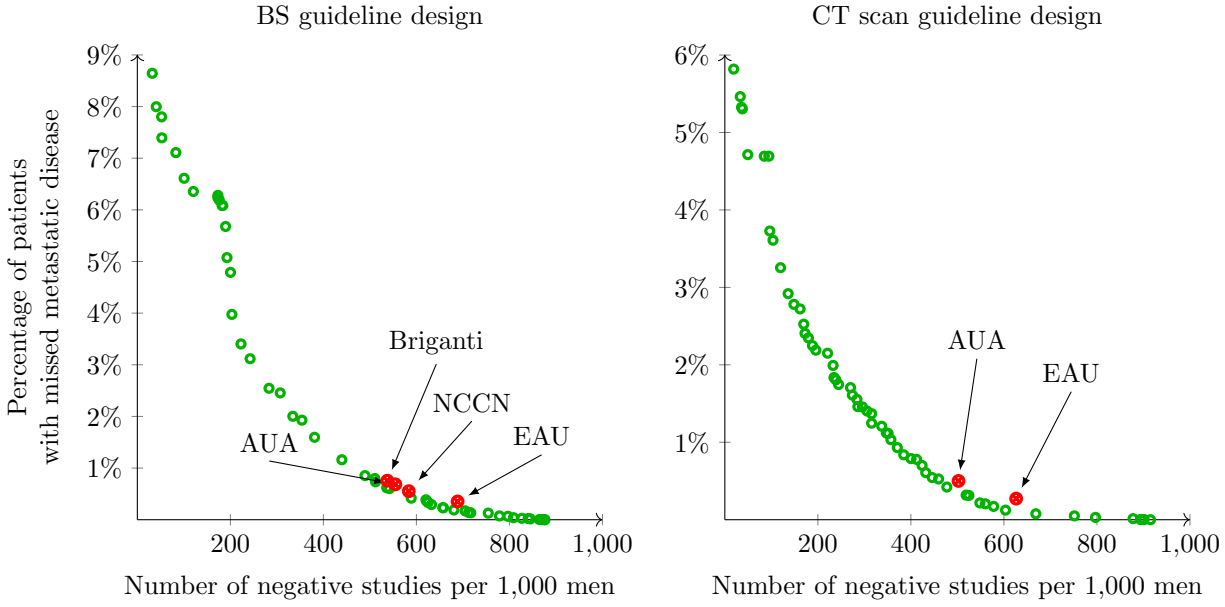

Figure 2.7: Trade-off curves for the BS and CT scan imaging guidelines with respect to the missed metastatic cancer rate and the number of negative studies estimated on the validation samples.

Additionally, we estimated the change in total number of imaging tests that can be expected from successful implementation of each clinical guideline compared to current practice (Table 2.10). After assessing the performance of the available clinical guidelines on the appropriate use of BS and CT scan in newly-diagnosed PCa patients, we showed that implementation of the AUA guidelines would reduce the total number of BS and CT scans by 25% and 26%, respectively, compared to current imaging practices; moreover, our

models predicted the percentage of patients with missed metastatic disease to be less than 1% [110, 133].

Table 2.10: Performance of the published guidelines for recommending staging BS and CT scan.

Clinical guidelines	No. of patients to be scanned	No. of metastases missed	No. of negative imaging studies	No. of patients to be scanned	Expected no. of metastases missed	Expected no. of negative imaging studies
Bone scan	**Patients with BS (n = 416)**			**Entire population (n = 1,519)**		
EAU [115]	288 (69.2)	1 (0.1)	127 (30.5)	405 (26.7)	10 (0.7)	350 (23.0)
AUA [161]	255 (61.3)	1 (0.1)	160 (38.5)	314 (20.7)	12 (0.8)	261 (17.2)
NCCN [117]	265 (63.7)	1 (0.1)	150 (36.1)	332 (21.9)	12 (0.8)	278 (18.3)
Briganti's CART [30]	244 (58.7)	5 (1.2)	167 (40.1)	292 (19.2)	13 (0.9)	243 (16.0)
CT scan	**Patients with CT scan (n = 643)**			**Entire population (n = 2,380)**		
EAU [74]	429 (66.7)	1 (0.2)	213 (33.1)	660 (27.7)	9 (0.4)	581 (24.4)
AUA [36]	355 (55.2)	2 (0.3)	286 (44.5)	475 (20.0)	11 (0.5)	399 (16.8)

The numbers in parentheses are the percentages. EAU: European Urological Association; AUA: American Urological Association; NCCN: National Comprehensive Cancer Network; CART: classification and regression tree.

2.5. Implementation and Impact

MUSIC is a physician-led, statewide quality-improvement collaborative that includes 43 urology practices in the state of Michigan and about 90% of the urologists in the state. A complete timeline of our project is shown in Figure 2.8. The first stage of the project was data collection. MUSIC has data abstractors at each MUSIC urology practice in the state to collect and verify the validity of the data in the MUSIC data repository. The next stage was model development, which included variable selection, model fitting, and guideline evaluation using the predictive models. During this stage, we had regular weekly meetings with the co-directors of MUSIC to update them with our results and to obtain feedback from a clinical perspective. The next stage was model validation, during which we performed both internal and external validation. We subsequently started the guideline design stage, during which our results for the performance of varying guidelines were presented to practicing urologists. Although risk-based guidelines performed well, MUSIC decided to endorse a threshold-based policy for several reasons: (1) according to our models these guidelines were near-optimal with respect to the miss rate and image usage; (2) a threshold-based policy is easier to understand and implement than a risk-based policy; and (3) similar guidelines had already been endorsed by the AUA.

Our results and the resulting proposed guidelines were first reviewed by the MUSIC

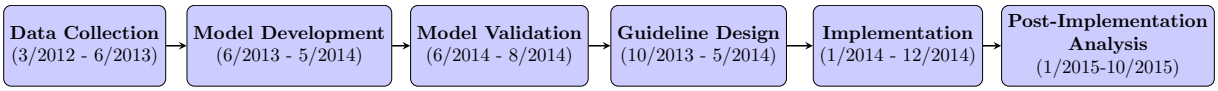

Figure 2.8: Project timeline from data collection to post-implementation analysis.

Imaging Appropriateness Committee, which included a sample of practicing urologists from across the state and a patient representative. Next, a selected subset of guidelines were reviewed at a MUSIC collaborative-wide meeting with approximately 40 urologists, nurses, and patient advocates. After achieving consensus with the collaborative, the MUSIC consortium instituted statewide, evidence-based criteria for BS and CT scan, known as the MUSIC Imaging Appropriateness Criteria (see the following Youtube video: `https://youtu.be/FEIxb_HRHAA`). The criteria recommends a BS for patients with PSA $>$ 20 ng/mL or Gleason score $\geq$ 8 and recommends a CT scan for patients with PSA $>$ 20 ng/mL, Gleason score $\geq$ 8, or clinical T stage $\geq$ cT3.

Recognizing the importance of clinical judgment in staging decisions, the MUSIC consortium set a statewide goal of performing imaging in $\geq$ 95% of patients that meet the criteria and in $<$ 10% of patients that do not meet the criteria. To implement the work, our collaborators presented our results at collaborative-wide meetings with "clinical champions", who returned to their practices to present the results to their own practice group. As part of this project, MUSIC members were provided with a toolkit including placards with the criteria (shown in Figure 2.9) and explanations for patients. After implementation, members also received comparative performance feedback that detailed how well their practice patterns correlated with the MUSIC Imaging Appropriateness Criteria.

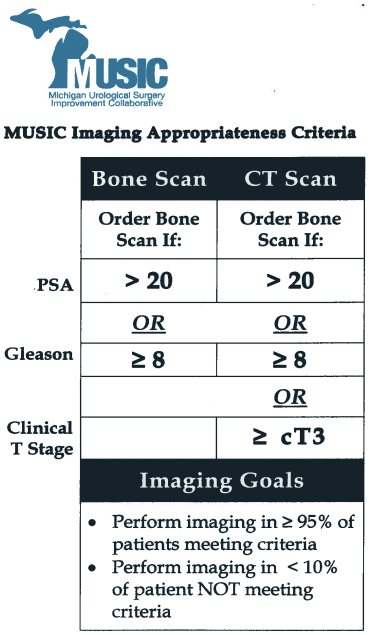

Figure 2.9: MUSIC placard.

After implementing this intervention in 2014, the MUSIC collaborative measured post-intervention outcomes from January to October 2015. The results showed an increase in the use of BS and CT scans in patients that meet the criteria from 82% to 84% and from 74% to 77%, respectively. Although these values are not $>$ 95%,

the MUSIC consortium has made measurable improvements in a short period of time and additional increases are anticipated. As shown in Figure 2.10, the MUSIC collaborative decreased the use of BS and CT scans in patients that do not fit the criteria from 11% to 6.3% and from 14.7% to 7.6%, respectively. Both of these values are below their goal of performing imaging in $< 10\%$ of patients that do not meet the criteria. These results were presented at the AUA Annual Meeting in San Diego, CA [82].

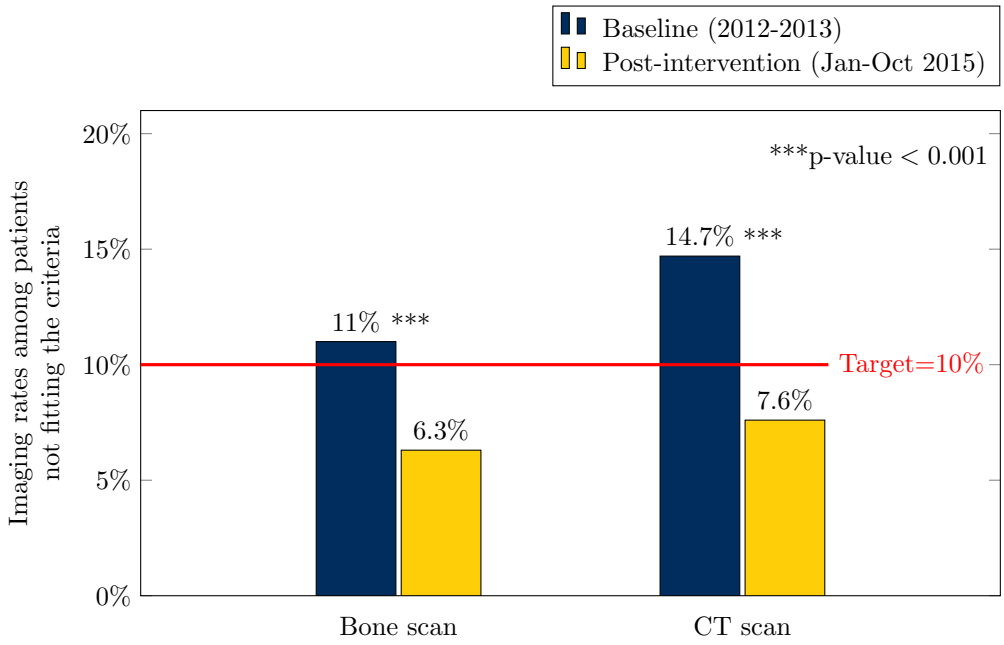

Figure 2.10: Avoidance of low-value imaging using MUSIC Criteria.

2.6. Conclusions

This work has had a significant societal impact by decreasing the chance of missing a case of metastatic cancer and substantially reducing the harm from unnecessary imaging studies. An important finding pertaining to the post-implementation is that the rate of positive BS or CT scan when indicated remained stable both at the baseline and after the intervention [82]. Additionally, this intervention has reduced healthcare costs without having a negative impact on patient outcomes. We have estimated that the MUSIC collaborative saved more than $\$262,000$ in 2015 through reducing unnecessary imaging studies and these savings will continue to accrue in future years. This is a conservative estimate of savings, because these are early results post-implementation that do not account for the savings from avoiding

unnecessary follow-up procedures for false-positive imaging studies. These savings also do not quantify the more important reduction in harm to patient health from reduced radiation exposure, fewer unnecessary follow-up procedures, and decreased patient anxiety.

The overuse of imaging in the staging of low-risk PCa patients was raised as the top priority by the AUA *Choosing Wisely* initiative. Our work extends this recommendation showing how patient data collected in a large region can be used to improve the prevision of clinical decision making. The publications of this work are building national recognition of this effort that may result in improvements beyond the state of Michigan [82, 110, 133]. Recently, our publications have been cited in the new National Comprehensive Cancer Network (NCCN) guidelines [117]. Thus, our work may ultimately influence national policy for cancer staging.

This work has paved the way for the development of guidelines based on individual risk factors in other areas; thus, we anticipate additional improvements to come in future years by building upon the successes described above. For example, this work has led to prototype of an iPhone app that reports a patient's risk of positive BS or CT scan, as well as a biopsy outcome prediction calculator, which has been implemented as a web-based decision support system called AskMUSIC (see `https://askmusic.med.umich.edu/`).

The multi-step approach presented in this chapter can be applied to other forms of cancer. For example, breast cancer is the most common malignancy in women, with approximately $250,000$ cases diagnosed yearly in the United States [1]. Imaging tests such as positron emission tomography (PET), CT scan, integrated PET with CT scan and BS are used to determine the extent of cancer. There are several clinical guidelines discouraging the routine use of staging scans for patients with clinically localized breast cancer [66, 142]. The framework presented in this chapter can be applied to determine the optimal imaging guidelines for breast cancer staging using observational data where there exists systematic bias in the selection of patients for imaging tests. Moreover, our framework allows the decision maker to choose from a spectrum of imaging guidelines including the published clinical guidelines and the classification rules to be generated via machine learning methods based on the trade-off between the available budget on imaging and the allowable rate of missed metastatic cases for patient population being considered.

Chapter 3.

Robust Optimal Design of Coordinated Imaging Protocols

3.1. Introduction

Imaging is playing an increasingly important role in the detection of cancer. In this context, the goal of imaging is accurate disease characterization through synthesis of anatomic, functional, and molecular imaging information. Many different kinds of imaging modalities such as bone scan (BS), computed tomography (CT scan), magnetic resonance imaging (MRI) and positron emission tomography (PET) are now being used for all facets of cancer — diagnosis and localization, staging, active surveillance and recurrence monitoring. With the recent changes in clinical care, scientific discoveries and technological advances, the spectrum of available options is continuously evolving. However, despite the tremendous advances in imaging in recent years, difficulty remains in selecting tests for patients as each imaging test has advantages, disadvantages, and specific indications.

In the selection of an imaging test, the benefits of imaging must be weighed against potential risks and harms. One of the risks of imaging tests is that they expose patients to radiation. Although technical innovations have helped reduce the radiation dose from imaging, the lifetime attributable risk of cancer from radiation exposure remains a significant concern [81]. Moreover, because of their imperfect nature, the available imaging tests can cause false positive and false negative results. The former leads to anxiety and unnecessary referral of patients for costly and invasive tests and treatments; the latter results in patients' cancer going undetected, and potentially progressing to a life threatening stage. Additionally, unnecessary imaging blocks access to the imaging resources for other

patients, and unnecessarily increases healthcare costs.

Cancer can spread regionally to lymph nodes, tissue or organs, and also to other parts of the body (i.e, metastasis). Metastatic cancer is associated with considerable morbidity, impaired health-related quality of life and reduced survival. Clinical staging is the process of determining the identification and extent to which cancer has spread in the body. Staging is important not only for prediction of prognosis, but also for providing a balanced approach in designing a treatment plan for individual patients diagnosed with cancer addressing the tradeoff between the benefits of treatments with the potential risks of complications and effects on quality of life. For example, patients at an early-stage of cancer may benefit from surgery or radiation treatment, while patients at a more advanced-stage may need to be treated with chemotherapy.

Imaging has become the mainstay for cancer staging as it can guide treatment selection, as well as treatment planning. In order for an imaging recommendation to be appropriate, it needs to fully address the perspectives of patients and physicians. From a patient and physician point of view, it needs to place an emphasis on individual health outcomes. From a health system perspective, it needs to weigh the benefits and harms of imaging at the population level. To satisfy both perspectives, the models we develop are aimed at guiding the allocation of imaging resources with the goal of minimizing imaging subject to constraints on the rate of missed disease at the patient population level. We provide evidence later in this chapter that this approach is consistent with both the patient and population perspective.

In this chapter, we evaluate the important evolving role of multi-modality imaging for detection and localization of cancer from an operations research perspective. We study this problem in the context of prostate cancer (PCa); however, the models and methods we describe could apply equally well to many other forms of cancer. To optimize the decision making for PCa imaging, we combine optimization and predictive analytics methods to develop models for predicting cancer outcomes of imaging tests, and to integrate these models into robust optimization models to design optimal imaging guidelines that can account for errors in the predictions. Given its clinical significance, our work generates important insights and findings for clinicians, health systems and other stakeholders seeking a satisfactory tradeoff between the benefits and harms of using these imaging tests for the staging of patients with newly-diagnosed PCa.

The remainder of this chapter is organized as follows. In Section 3.2, we provide background on PCa staging, predictive modeling, diagnostic testing and robust optimization.

In Section 3.3, we introduce mathematical notation and provide mathematical formulations of the decision problem in the cases of a single imaging test and multiple imaging tests. We also analyze the structural properties of the problems, and provide robust formulations as well as heuristic algorithms. In Section 3.4, we describe the methodological approach for development and validation of a multinomial logistic regression model, and the analytical approach to quantify statistical variation in probability estimates obtained from predictive models. In Section 3.5, we present results using real medical data and discuss our findings. Finally, in Section 3.6, we conclude with a summary of our findings.

3.2. Background and Literature Review

This section provides background on (1) PCa staging, (2) diagnostic testing decisions, (3) predictive modeling in medicine, and (4) robust optimization. Furthermore, we review the most relevant literature on each of these topics.

3.2.1. Prostate Cancer Staging

The goal of PCa staging is to determine whether the cancer has metastasized to lymph nodes or to bones. PCa is a solid tumor that exhibits a tendency to metastasize to the bones. The skeleton is the site of first and main metastasis in about 80% patients with PCa; therefore, bone metastases are one of the most important prognostic factors [165]. Bone metastases are associated with considerable morbidity (pain, reduced mobility, pathological fractures, spinal cord and nerve compression), reduced survival (5-year survival is 3%) and also significant health economic implications including the costs of systematic therapies, imaging and hospital admissions [31, 127, 178]. The presence of lymph node metastasis is also an important prognostic factor, indicating great risk for progression to bone metastases and death [63]. As of today, metastatic PCa is still considered incurable but there are treatment options that can increase survival. Therefore, accurate staging is crucial for the clinical management of PCa changing from possible cure to alleviating symptoms and improving quality of life.

Conventional imaging tests for PCa staging include BS and CT scan for detection of bone and lymph node metastases, respectively. However, not all men with newly-diagnosed PCa are at the same risk of harboring metastatic cancer. In screening trials, bone metastases are detected at diagnosis in less than 10% of the patients [165], and lymph node metastases

in between 4% to 6% of the patients [28]. These summary statistics suggest that it may not be necessary to perform the imaging tests for every new patient. This is an important consideration because there are harms associated with both under- and over-imaging. Under-imaging results in patients' metastatic PCa going undetected. In such cases, patients are subjected to treatment, such as radical prostatectomy (surgical removal of the prostate), that is unlikely to benefit the patient, and can lead to serious side effects and negative health outcomes due to delays in chemotherapy. Over-imaging causes potentially harmful radiation exposure [96, 128, 149], anxiety for the patient, and false positive findings that lead to risky and painful follow-up procedures (i.e., bone biopsy). Not only do these imaging tests expose the patient to excess radiation, but they also increase financial and time burdens both on the patient and healthcare system.

To facilitate the optimal imaging of newly-diagnosed PCa, professional societies such as the National Comprehensive Cancer Network (NCCN), American Urology Association (AUA), American Cancer Society (ACS) and European Association of Urology (EAU) have established international evidence-based guidelines indicating the need for BS and CT scan only in patients with certain unfavorable risk factors; however, the guidelines vary in their recommendations [30, 36, 74, 115, 117, 161]. Thus, there exists persistent variation in utilization of these imaging tests among urologists, including unnecessary imaging in patients at low risk for metastatic disease and potentially incomplete staging of patients at high risk. In 2012, the AUA highlighted the need to reduce imaging for low-risk PCa in the *Choosing Wisely* campaign, a multidisciplinary effort to reduce unnecessary imaging, decrease overuse of healthcare resources, and improve quality of care [6].

In Chapter 2, we studied the imaging problem independently for BS and CT scan in a population-based sample of men with newly-diagnosed PCa from the diverse academic and community practices in the Michigan Urological Surgery Improvement Collaborative (MUSIC), which includes 90% of the urologists in the state [110, 133]. We used data-analytics approaches to develop and validate risk prediction models to help urologists make PCa staging decisions. In addition to the international evidence-based guidelines, we implemented classification modeling techniques to develop accurate decision rules recommending patients for imaging tests. These models were used to design guidelines that weigh the benefits and harms of radiological imaging. MUSIC implemented the proposed guidelines which miss less than 1% of metastatic cancers while reducing unnecessary imaging by more than 40%. In the work presented in this chapter, we seek to improve efficiency and effectiveness of imaging by coordinating multiple imaging tests by drawing on the

predictive models developing in Chapter 2.

3.2.2. Diagnostic Testing Decisions

In this chapter, we study robust optimization models for coordinated imaging protocols for PCa staging that consider different combinations of BS and CT scan. There is no evidence-based imaging guideline addressing the need for both BS and CT scan in a holistic approach like this: clinicians often order both imaging tests simultaneously or no tests. However, given the correlation observed between BS and CT scan results, the result of one imaging test can be used to predict the result of another follow-on test, which in turn, motivates a sequential imaging paradigm in which some patients may benefit from having the imaging tests one at a time. In a more general context, applicable also to diseases other than PCa staging, we are concerned with the problem of optimal allocation of *composite* diagnostic tests that may combine multiple tests to more accurately and efficiently detect the presence of disease. This is an important problem as imaging resources are expensive and limited. Thus, poor decisions can lead to serious health outcomes, resulting in high healthcare costs and a significant reduction in quality of life.

The optimal selection of diagnostic tests for disease screening has been studied in the context of blood screening where the goal is to reduce the risk of transfusion-transmitted infectious diseases (TTIs), including the human immunodeficiency virus, hepatitis viruses B and C, human T-cell lymphotropic virus and syphilis [25, 26, 177]. In blood screening, the testing strategies need to consider that (1) the screening tests are not perfectly accurate, (2) there are often multiple tests available for each TTI, and (3) most tests are expensive and resources are limited. Earlier work addressing different aspects of this problem showed that optimized screening strategies result in a more effective and efficient screening for donated blood compared to the current screening strategies, without increasing resource requirement.

Recently, El-Amine et al. [5] expanded previous research on the optimization for blood screening by accounting for the uncertainty in the prevalence rates of TTIs and the limited information that the decision maker has. In the author's proposed optimization framework, nonlinear continuous knapsack problems are utilized, and robust blood screening strategies are shown to offer substantial reduction in the risk of TTIs compared to the screening strategies compliant with the Food and Drug Administration (FDA), at limited levels of budget available to be allocated for screening.

The literature on the optimal test selection problem for blood screening illustrates the benefits of an optimization-based framework in resource-constrained settings. In this chapter, similar to the existing literature for optimized blood screening, we are interested in the decision of how to optimally assign *patient types* (to be discussed in Section 3.3) to imaging protocols to minimize the burden of imaging while ensuring that the rate of missed metastatic cancer cases is below a certain threshold (e.g., budget) in the population. A commonly used formulation that is also related to our problem is the traditional knapsack problem, which selects, from a set of candidates, each with a known reward and cost, an optimal set that is budget-feasible and that maximizes the total reward of the selected items.

The work presented in this chapter differs from the existing literature on diagnostic testing decisions for blood screening in two main ways. First, we present a new multiple choice knapsack problem (MCKP) formulation in which the objective is to minimize imaging in the population while generating budget-feasible assignments of patient types into coordinated imaging protocols. In contrast to our objective, the previous work on blood screening aims to find the optimal test selection that achieves a low TTI risk. Second, our model involves the use of predictive models for estimating the probability of imaging outcomes based on patients' risk factors, and the individual probability estimates obtained from predictive models are used to define uncertainty sets for the model parameters that depend on predictions. The closest work that considers parameter uncertainty in this framework is given in the work by El-Amine et al. [5]. It assumes that the only information available to the decision maker is the support of the random prevalence rate vector of infectious diseases, and this information relies on the estimates reported in the literature in other studies. To our knowledge, we are the first to integrate predictive models into optimization models to determine robust imaging protocols that take errors in predictions into account.

Consistent with the incentive of published imaging guidelines to reduce the overuse of imaging at staging, the objective of the optimization-based models we develop is to reduce the total number of imaging tests performed at the population-level, subject to a certain budget level. In this context, the budget represents the maximum acceptable rate of missed metastatic disease in the population. The significant impact of the preoperative detection of metastases on the selection of appropriate treatment, quality of life and survival underscores the importance of incorporating the missed disease rate as a constraint into our mathematical formulations.

In our proposed framework, we take a clinical perspective that focuses on patients' health

but not specifically on healthcare costs. The reasons for this are two-fold: (1) more than 90% of PCa patients are diagnosed with clinically localized disease (i.e., tumor is confined within prostate), rendering the routine use of radiographic staging largely unnecessary [84] and (2) in that small high-risk population; however, it is critical to precisely rule out the presence of metastasis as it is associated with significant symptom burden and increased mortality [31, 32]. Nevertheless, alternative model formulations that incorporate cost or other criteria are easily developed.

The appropriateness of testing is dependent on the likelihood that a patient has the suspected disease, which in turn may depend on a number of clinical and demographic factors. Hence, we study the problem from a perspective in which individualized patient probability estimates for the presence of metastatic disease are estimated using predictive models. In the following subsection, we briefly discuss predictive modeling in medicine, and review the properties of a number of most commonly used measures for the assessment of the performance of a predictive model.

3.2.3. Predictive Modeling in Medicine

Predictive models are increasingly used to provide guidelines for clinical decision making and the personalized management of diseases [48, 90, 114, 153]. For a predictive model to be useful in decision making, it must provide validated and accurate estimates of probabilities of specific health conditions or outcomes. As discussed in Chapter 2, validating a predictive model using the development sample will introduce bias, known as optimism, because the model will typically fit the training dataset better than a new dataset. Therefore, it is essential to quantify any optimism in the predictive performance of the developed model. Internal validation techniques include randomly splitting the data, bootstrapping or cross-validation [112, 113, 154].

Because we incorporate the probability estimates into the optimization models to determine the optimal allocation of imaging protocols for the population, obtaining well-calibrated probability estimates is of more importance than the discriminate value. A predictive model is well-calibrated if, for example, it produces a predicted probability of 40% risk of having the disease for one patient, and similar patients would truly be diagnosed with the disease 40% of the time. Poor calibration, on the other hand, will lead to systematic errors in the model performance.

Similar to Chapter 2, the discrimination of the predictive models is quantified using

the area under the ROC curve (AUC), and the calibration of the predicted probabilities is assessed via the Brier score. We also assess the model calibration graphically with calibration plots that were described in Chapter 2. Figure 2.3 in Chapter 2 shows the calibration plots for the multivariate logistic regression models developed for BS and CT scan [108].

In practice, it is always the case that a predictive model will have imperfect calibration. Several factors such as the challenges in data collection and management (incomplete, heterogeneous, incorrect, or inconsistent data), small sample size, existence of large numbers of candidate predictors and the increased uncertainty surrounding rare events contribute to the imperfect nature of predictive models. In our proposed framework, the predictions are used to inform the assignment of imaging protocols to patients on the basis of their estimated probability of disease. Therefore, it is important to immunize imaging decisions against the *statistical errors* in calibration. For this purpose, we utilize robust optimization concepts and techniques that are tailored to incorporate this error, and attempt to identify solutions that are robust with respect to the implied parameter uncertainty. In the following subsection, we provide background on robust optimization and a review of the relevant literature.

3.2.4. Robust Optimization

Traditional modeling approach for decision-making assumes perfect information, i.e., the input data is precisely known and equal to some nominal value. However, in real-world applications data is often incomplete or contains errors. The data errors can be derived from measurement or estimation errors resulting from the lack of knowledge of the model parameters. For example, as illustrated in Figure 2.3, the statistical errors in our context result from the insufficiency of the predictive models to provide perfectly calibrated estimates of diagnostic probabilities. As illustrated by [18], solutions to optimization models can exhibit substantial sensitivity to errors in the parameters of the mathematical models, therefore rendering a computed solution infeasible, suboptimal, or both.

There are two common approaches to handle data uncertainty in optimization: stochastic and robust optimization. In the stochastic optimization approach, the main assumption is that the true probability distributions of the random variables are known or can be estimated. For details on stochastic optimization, we refer to [24, 138] and [129] for an overview of solution techniques. Determining an appropriate probability density function

based on historical data can be difficult in some cases. Moreover, in some situations the lack of historical data makes it impossible to obtain an accurate probabilistic description of the uncertainty. In the robust optimization approach, the uncertainty model is not stochastic, but deterministic and set-based, i.e., the value of an uncertain parameter varies in a prespecified *uncertainty set*. Thus, robust optimization seeks to mitigate sensitivity of the model-based solution to variations in the model parameters.

Robust optimization has emerged as a powerful modeling tool to handle erroneous or noisy data in decision-making over the last decade because of its computational tractability and practicability. For a detailed overview, see [16, 20] and [21], and the references therein. In contrast to stochastic optimization where the goal is to optimize an expectation, the goal in the robust optimization framework is to find a solution that is feasible for any realization of the uncertainty in a given uncertainty set. In other words, it optimizes against the worst-case instances using a min-max objective.

Soyster's approach is the earliest work in this area [150]. He proposed a linear optimization problem in which each coefficient can vary independently in an interval. Although the proposed approach guarantees feasible solutions with respect to every realization of the coefficients, the resulting model produces overly conservative solutions as the uncertain coefficients can take their worst values simultaneously. To avoid over-conservatism, [17, 19, 55] and [56] independently developed robust models of which the uncertainty sets for the data are ellipsoids, and proposed efficient algorithms to solve convex optimization problems under data uncertainty. The tradeoff between robustness and performance is controlled by the decision-maker through the size of the ellipsoidal sets. A drawback of the proposed robust models with ellipsoidal uncertainty sets is that it increases the complexity of the problem, e.g., the resulting robust counterpart formulations involve quadratic constraints leading to second-order cone programming.

Other prominent studies that attempt to avoid over-conservatism include the works of [22, 23]. They proposed a robust optimization approach based on polyhedral uncertainty sets. In their proposed framework, the *budget of uncertainty* can be adjusted by controlling the number of uncertain parameters that are allowed to deviate from their nominal values, thus providing a way of incorporating different attitudes of the decision-maker toward risk (e.g., risk-averse, risk neutral, or risk-seeking). Moreover, as the proposed models retain a linear structure similar to the framework of [150], they have the advantage of being less demanding computationally compared with the former robust models with ellipsoidal sets. All these factors have enabled this approach to be widely used in different areas

including logistics and production systems, portfolio selection problem, data envelopment analysis [83]. This robust counterpart optimization approach is also shown to be directly applied to discrete optimization models [22]. Given the practicability and tractability of the budgeted robust counterpart optimization approach in combinatorial problems that are subject to data uncertainty, we adopted this approach in our mathematical formulations for the robust optimal design of imaging protocols.

The robust counterpart of an uncertain optimization model is a deterministic formulation in which model parameters are assumed to be uncertain, but symmetrically distributed over a bounded interval, e.g., the uncertainty set U. The structure and scale of the set U is determined by the decision maker. The structure refers to the geometry of the constraint set U. The two most common ways of defining the geometry of uncertainty sets are ellipsoidal and polyhedral sets, which will be the main focus of the review in this section. The scale refers to the magnitude of the deviations of the uncertain parameters from their nominal values.

In Soyster's model, the worst-case solution is guaranteed to be feasible for all realizations of the uncertain parameters. Bertsimas and Sim [23] relaxed this condition by proposing that not every parameter will take its boundary value at the optimal solution. To represent the robust counterpart with budgeted uncertainty in mathematical terms, we start by considering the following nominal linear program (LP) model:

$$\begin{aligned} \text{maximize} \quad & \mathbf{c}'\mathbf{x} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} \leq \mathbf{b} \\ & \mathbf{l} \leq \mathbf{x} \leq \mathbf{u} \end{aligned} \tag{3.1}$$

where $\mathbf{c}, \mathbf{l}, \mathbf{u}$ are n-vectors, $\mathbf{A}$ is an $m \times n$ matrix, and $\mathbf{b}$ is an m-vector. Without loss of generality, it is often assumed that the data uncertainty only affects the elements in matrix $\mathbf{A}$. If the coefficients of the objective function are subject to uncertainty, we can then maximize the objective z, and add the constraint $z - \mathbf{c}'\mathbf{x}$ into $\mathbf{A}\mathbf{x} \leq \mathbf{b}$ [23]. Let J_i represent the set of coefficients in row i of the matrix $\mathbf{A}$ that are subject to uncertainty. Each entry a_{ij} is modeled as a symmetric and bounded random variable $\tilde{a}_{ij}$, representing the actual value of the coefficient, and can take values in the range $[a_{ij} - \hat{a}_{ij}, a_{ij} + \hat{a}_{ij}]$, where a_{ij} is the nominal value of $\tilde{a}_{ij}$, and $\hat{a}_{ij}$ is the maximum positive perturbation of the

corresponding uncertain coefficient. Thus, we can define $\tilde{a}_{ij}$ as:

$$\tilde{a}_{ij} = a_{ij} + \eta_{ij}\hat{a}_{ij} \tag{3.2}$$

where η_{ij} is a random variable with an unknown but symmetric distribution and takes values in the range $[-1, 1]$. As noted in [23], it is unlikely that all of the a_{ij}, $j \in J_i$, will change. For this purpose, the following polyhedral uncertainty set for the matrix $\mathbf{A}$ is proposed:

$$U^A = \left\{ \tilde{a}_{ij} = a_{ij} + \eta_{ij}\hat{a}_{ij} \;\middle|\; \sum_{j=1}^{n} |\eta_{ij}| \leq \mathbf{\Gamma}_i, \forall i \in J_i \right\} \tag{3.3}$$

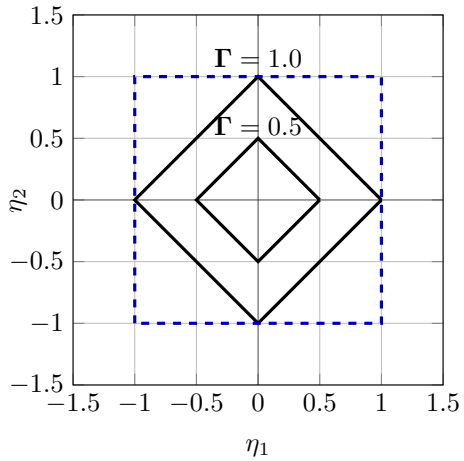

Figure 3.1: Polyhedral uncertainty set.

Polyhedral uncertainty associated with two different values of $\mathbf{\Gamma}_i$, for a problem of two uncertain coefficients is shown in Figure 3.1. If $\mathbf{\Gamma}_i$ is integer, a solution is considered as a robust solution that is protected against all cases in which at most $\mathbf{\Gamma}_i$ coefficients of the ith constraint are allowed to vary under the proposed polyhedral uncertainty set in (3.3).

The corresponding robust counterpart is defined as follows:

$$
\begin{aligned}
&\text{maximize} \quad \mathbf{c}'\mathbf{x} \\
&\text{subject to} \\
&\sum_{j=1}^{n} a_{ij}x_j + \max_{\{S_i \cup \{t_i\} \mid S_i \subseteq J_i, |S_i| = \lfloor \mathbf{\Gamma}_i \rfloor, t_i \in J_i \setminus S_i\}} \left\{ \sum_{j \in S_i} \hat{a}_{ij} y_j + (\mathbf{\Gamma}_i - \lfloor \mathbf{\Gamma}_i \rfloor)\hat{a}_{it_i} y_{t_i} \right\} \leq b_i, \quad \forall i \\
&-y_j \leq x_j \leq y_j, \quad \forall j \\
&l_j \leq x_j \leq u_j, \quad \forall j \\
&y_j \geq 0, \quad \forall j
\end{aligned}
\tag{3.4}
$$

For every constraint i, a parameter $\mathbf{\Gamma}_i$ is introduced that takes value in the interval $[0, |J_i|]$. The solution of the model in (3.4) is protected against all cases that up to $\lfloor \mathbf{\Gamma}_i \rfloor$ of coefficients are allowed to change and one coefficient, a_{it_i}, changes by $(\mathbf{\Gamma}_i - \lfloor \mathbf{\Gamma}_i \rfloor)\hat{a}_{it_i}$. By strong duality, Bertsimas and Sim [23] proved that the robust counterpart in (3.4) is equivalent to the following LP model:

$$
\begin{aligned}
&\text{maximize} \quad \mathbf{c}'\mathbf{x} \\
&\text{subject to} \\
&\sum_{j=1}^{n} a_{ij}x_j + z_i\mathbf{\Gamma}_i + \sum_{j \in J_i} p_{ij} \leq b_i, \quad \forall i \\
&z_i + p_{ij} \geq \hat{a}_{ij} y_j, \quad \forall i, j \in J_i \\
&-y_j \leq x_j \leq y_j, \quad \forall j \\
&p_{ij} \geq 0, \quad \forall i, j \in J_i \\
&y_j \geq 0, \quad \forall j \\
&z_i \geq 0, \quad \forall i
\end{aligned}
\tag{3.5}
$$

The parameter $\mathbf{\Gamma}_i$ is referred to as the budget of uncertainty of constraint i. It allows the decision maker to balance the tradeoff between the protection level of the constraint and the level of conservatism of the solution. If $\mathbf{\Gamma}_i = 0$, the η_{ij}'s are forced to 0 so that the coefficients $\tilde{a}_{ij}$ equal to their nominal values a_{ij}, which implies that there is no protection against uncertainty. If $\mathbf{\Gamma}_i = |J_i|$, the constraint $\sum_{j=1}^{n} |\eta_{ij}| \leq \mathbf{\Gamma}_i$ becomes redundant as $|\eta_{ij}| \leq 1$ for all $i \in J_i$, which implies that the ith constraint is fully protected against uncertainty yielding a very conservative solution. Thus, as $\mathbf{\Gamma}$ increases, more protection is given and the solution is more robust to uncertainty.

We now briefly describe how the robust optimization framework based on polyhedral

uncertainty sets can be extended to discrete robust optimization problems. Consider a nominal mixed integer program (MIP) that is the nominal LP model with the additional constraint that the first k of a set n variables are integrals. The uncertainty for the constraint matrix $\mathbf{A}$ is modeled in the same way as described for the continuous case. The uncertainty for the cost vector $\mathbf{c}$ is modeled such that each entry c_j takes values in the range $[c_j, c_j + d_j]$, where d_j is the estimate of the deviation from the nominal cost coefficient c_j, $d_j \geq 0$ for all j. Similar to (3.4), the robust counterpart for this uncertain MIP is defined as follows:

$$
\begin{aligned}
&\text{maximize} \quad \mathbf{c}'\mathbf{x} + \max_{\{S_0 \mid S_0 \subseteq J_0, |S_0| \leq \mathbf{\Gamma}_0\}} \left\{ \sum_{j \in S_0} d_j y_j \right\} \\
&\text{subject to} \\
&\sum_{j=1}^{n} a_{ij} x_j + \max_{\{S_i \cup \{t_i\} \mid S_i \subseteq J_i, |S_i| = \lfloor \mathbf{\Gamma}_i \rfloor, t_i \in J_i \setminus S_i\}} \left\{ \sum_{j \in S_i} \hat{a}_{ij} y_j + (\mathbf{\Gamma}_i - \lfloor \mathbf{\Gamma}_i \rfloor) \hat{a}_{it_i} y_{t_i} \right\} \leq b_i, \quad \forall i \\
&-y_j \leq x_j \leq y_j, \quad \forall j \\
&l_j \leq x_j \leq u_j, \quad \forall j \\
&y_j \geq 0, \quad \forall j \\
&x_i \in \mathbb{Z}, \quad \forall i = 1, \ldots, k
\end{aligned}
\tag{3.6}
$$

Furthermore, Bertsimas and Sim [22] proved that the robust counterpart in (3.6) has the following equivalent MIP formulation:

$$
\begin{aligned}
&\text{minimize} \quad \mathbf{c}'\mathbf{x} + z_0 \mathbf{\Gamma}_0 + \sum_{j \in J_0} p_{0j} \\
&\text{subject to} \quad \sum_{j=1}^{n} a_{ij} x_j + z_i \mathbf{\Gamma}_i + \sum_{j \in J_i} p_{ij} \leq b_i, \quad \forall i \\
&\qquad z_0 + p_{0j} \geq d_j y_j, \quad \forall j \in J_0 \\
&\qquad z_i + p_{1j} \geq \hat{a}_{ij} y_j, \quad \forall i \neq 0, j \in J_i \\
&\qquad p_{ij} \geq 0, \quad \forall i, j \in J_i \\
&\qquad y_j \geq 0, \quad \forall j \\
&\qquad z_j \geq 0, \quad \forall j \\
&\qquad -y_j \leq x_j \leq y_j, \quad \forall j \\
&\qquad l_j \leq x_j \leq u_j, \quad \forall j \\
&\qquad x_i \in \mathbb{Z}, \quad \forall i = 1, \ldots, k
\end{aligned}
\tag{3.7}
$$

3.3. Model Formulations and Analysis

In this section, we give a detailed description of how we utilized the robust optimization approach discussed in the previous section to design robust imaging guidelines for PCa staging. We start by considering the case of a single imaging test, and introduce some structural properties and algorithmic ideas which will later be used to establish a conceptual basis for the case of multiple tests. We tailor the notation introduced in Section 3.2.4 to the particular optimization problem at hand and point out the differences in the model formulations.

We introduce common notation to be used in the mathematical formulations for the two cases of the problem. We let T_1 and T_2 denote BS and CT scan, respectively. We consider the assignment of *patient types* into ideal imaging protocols and assume N types of patients differentiated on the basis of clinical risk factors, indexed by $j = 1, \ldots, N$. The most straightforward approach to define patient types is to use the risk factors that are highly associated with the presence of disease. Similar to treating a continuous variable as a dichotomous variable in statistical modeling, some established criterion or cutoff point can be used to create certain categories of risk factors that are clinically relevant. These categories can then be used to define patient types. We let w_j denote the proportion of patient type j in the population and D_j^k denote the random outcome of test T_k for patient type j and $k = 1, 2$, where $D_j^k \in \{-1, 1\}$: 1 corresponds to a positive test and -1 to a negative test.

We let $g_j(\cdot)$ denote the probability of an imaging outcome for patient type j, e.g., $g_j(D_j^1 = 1)$ and $g_j(D_j^2 = 1)$ denote the probability of a positive BS and CT scan, respectively. To estimate these patient-type specific probabilities, we used the predicted probability estimates for each patient in the study population obtained from the predictive models, and averaged them over each patient type to obtain the mean predicted probabilities for each type (discussed in detail in Section 3.4). We let m_j represent the expected proportion of a patient type j with missed disease, which we let refer to as the missed rate for the patient type j, under an imaging protocol. We let n_j represent the expected cost of imaging for a patient type j under an imaging protocol. Both m_j and n_j have an additional subscript, p, in the case in which there are multiple imaging protocols. We let α, $\alpha \in [0, 1]$, represent the maximum allowable rate of missed metastatic disease for the population determined by the decision maker, which we let refer to as the missed-rate budget.

3.3.1. Single Imaging Test Case

In the case of a single imaging test, we introduce a binary variable x_j defined as:

$$x_j = \begin{cases} 1, & \text{if patient type } j \text{ is assigned to imaging protocol} \\ 0, & \text{otherwise} \end{cases}$$

The missed rate for patient type j is defined as $m_j = w_j g_j(D_j^k = 1)$. In the presence of a single test, the expected cost of imaging becomes a constant in the optimization problems, and therefore the cost can be ignored, i.e., the expected cost of imaging for patient type j is defined as $n_j = w_j$ for Test k, which represents the expected number of imaging test performed for type j. The optimal assignment of patient types for imaging can be formulated as follows:

$$\min \left\{ \sum_{j=1}^{N} w_j x_j \;\middle|\; \sum_{j=1}^{N} m_j(1 - x_j) \leq \alpha,\ x_j \in \{0,1\}, \forall j \right\} \tag{3.8}$$

This model can be transformed into the standard knapsack formulation by setting $y_j = 1 - x_j$:

$$\max \left\{ \sum_{j=1}^{N} w_j y_j \;\middle|\; \sum_{j=1}^{N} m_j y_j \leq \alpha,\ y_j \in \{0,1\}, \forall j \right\} \tag{SIM}$$

which we refer to in our context as the *single imaging model* (SIM). Intuitively, SIM optimizes the selection of patient types to maximize the proportion of the population not imaged while not exceeding a given threshold value α of the missed disease rate.

In order for an imaging recommendation to be recognized as reasonable and fair by clinicians, it should be consistent with respect to each patient type's estimated probability of disease.

Definition 3.1. *A collection of decisions guiding the assignment of patient types to imaging is referred to as a consistent risk-ordering if it prescribes imaging for patient types of higher risk disease at least as much as for patient types at lower risk of disease.*

Unfortunately, the optimal SIM solution may not always be consistent with the risk ordering. To illustrate, consider a simple example with two patient types, $j = 1, 2$, and with $w_1 \gg w_2$ and $g_1(D_1 = 1) \ll g_2(D_1 = 1)$ such that $m_1 \leq \alpha$ and $m_2 \leq \alpha$ but $m_1 + m_2 > \alpha$. It follows that the greedy heuristic would image patient type 2 but the

optimal solution would be to image type 1. To address this problem, we propose greedy algorithms for the single imaging test case that provide near optimal solutions and are consistent risk-ordering.

We now show that although the optimal solution to SIM does not guarantee a consistent risk-ordering, we can obtain imaging recommendations that satisfy this condition using a simple greedy heuristic.

Proposition 3.1. *The SIM-Greedy algorithm generates a feasible solution with a consistent risk ordering that deviates from the optimal solution to SIM by at most* $w_{\max} = \max\{w_j \mid j = 1, \ldots, N\}$.

Proof. We let z^* denote the optimal value of SIM, and z^G the value obtained with the greedy algorithm. The well-studied greedy algorithm for the 0-1 knapsack problems orders the items (patient types) by their *efficiency*, which in this context corresponds to $w_j/m_j = 1/g_j$, in decreasing order, which is a consistent risk-ordering by definition. The bound on $z^* - z^G$ follows from a proof based on the LP-relaxation of the 0-1 knapsack problem (see p.19 in [86]). □

Proposition 3.1 shows that the gap between the optimal solution value, z^*, and the greedy solution value, z^G, depends on the characterization of the patient types. It should also be noted that all the data are assumed to be integer valued in the proof of this bound. Therefore, the magnitude of the transformed input values can negatively affect how far z^G can deviate from z^*.

Another important property of the SIM-Greedy algorithm is related to the robustness of the greedy solution. The greedy algorithm selects patient types for no imaging sequentially from low to high risk. As we will show in Section 3.5.2, the deviations in predictions tend to increase with respect to increases in g_j. Therefore, the greedy algorithm tends to provide a solution that is *immunized* against the uncertainty in the probability estimation, meaning the solution mitigates the influence of uncertainties in the allocation of imaging resources for the robust model described in the next subsection.

In addition to the SIM-Greedy algorithm, we propose a second algorithm, which we refer to as the *individualized* SIM-Greedy algorithm, that is motivated by the perspective of our clinical collaborators. This algorithm assigns individual patient types to imaging on the basis of their estimated probability of missed metastatic disease. Each type is assigned to imaging if the probability of missed disease for that type is above the budget α. It

Algorithm 3: SIM-Greedy.

$\overline{m} = 0$; $\overline{m}$ *is the total missed rate of the currently unimaged patient types*
$z^G = 1$; z^G *is the total imaging tests performed based on the current allocation*
Sort the patient types by their efficiency in decreasing order, i.e., $1/g_1 \geq 1/g_2 \geq \ldots \geq 1/g_N$.
for $j = 1$ **to** N **do**
  if $\overline{m} + m_j \leq \alpha$ **then**
    $y_j = 1$;
    $z^G = z^G - w_j$;
    $\overline{m} = \overline{m} + m_j$;
  else
    $y_j = 0$;
  end
end
Return the solution vector $(y_1, \ldots, y_N)$ and z^G.

is guaranteed that the greedy solution is feasible to SIM given the properties of w_j, i.e., $w_j \in [0, 1]$ for all j and $\sum_{j=1}^{N} w_j = 1$.

Algorithm 4: Individualized SIM-Greedy.

$z^G = 1$
for $j = 1$ **to** N **do**
  if $g_j(D_j^k = 1) \leq \alpha$ **then**
    $y_j = 1$;
    $z^G = z^G - w_j$;
  else
    $y_j = 0$;
  end
end
Return the solution $\mathbf{y}$ with value z^G.

We compare the results of SIM-Greedy and individualized SIM-Greedy to optimal solutions in Section 3.5.2.

The Robust Model

To account for the variations in the probability estimates obtained from predictive models, we adopt the uncertainty set proposed by [22]. Since the missed disease rates depend on

the probability estimates, which in turn are affected by statistical variation, we represent the uncertain m_j using the following uncertainty set:

$$U = \left\{ (\tilde{m}_j) \;\middle|\; \tilde{m}_j \in [m_j - \delta_j, m_j + \delta_j],\ \forall j;\ \sum_{j=1}^{N} \frac{|\tilde{m}_j - m_j|}{\delta_j} \leq \mathbf{\Gamma} \right\}$$

In this representation, the uncertain missed rate $\tilde{m}_j$ of type j has a nominal value m_j and a maximum variation δ_j ($\delta_j \geq 0$). The budget parameter $\mathbf{\Gamma}$ ($\mathbf{\Gamma} \in [0, |J|]$) is introduced to control the degree of solution conservatism as discussed in Section 3.2.4. Specifically, this uncertainty representation postulates that a missed rate $\tilde{m}_j \in [m_j - \delta_j, m_j + \delta_j]$ is determined for each type and the total variation in missed rates is less than or equal to $\mathbf{\Gamma}$.

Remark 3.1. *The formulation proposed by Bertsimas and Sim [23] assumes that $\tilde{m}_j$ is a symmetric random variable; however, in our application this is not necessarily true. This is still a reasonable assumption because the adversary seeks to maximize $\tilde{m}_j$, and in practice the lower bound is not achieved. Furthermore, because both the probability of missed metastatic disease and the proportion are less than* 1*, the upper bound on $\tilde{m}_j$ does not exceed* 1*.*

The robust counterpart formulation of SIM under the uncertainty set U is defined as follows:

$$\begin{aligned} \text{maximize} \quad & \sum_{j=1}^{N} w_j y_j \\ \text{subject to} \quad & \sum_{j=1}^{N} m_j y_j + \max_{\{S \cup \{t\} \mid S \subseteq J, |S| = \lfloor \mathbf{\Gamma} \rfloor, t \in J \setminus S\}} \left\{ \sum_{i \in S} \delta_i y_i + (\mathbf{\Gamma} - \lfloor \mathbf{\Gamma} \rfloor) \delta_t y_t \right\} \leq \alpha \\ & y_j \in \{0, 1\} \quad \forall j \end{aligned} \tag{3.9}$$

where J is the set of coefficients in the missed-rate budget constraint that are uncertain.

Problem (3.9) has the following equivalent MIP formulation:

$$\begin{aligned}
\text{maximize} \quad & \sum_{j=1}^{N} w_j y_j \\
\text{subject to} \quad & \sum_{j=1}^{N} m_j y_j + t\mathbf{\Gamma} + \sum_{j \in J} v_j \leq \alpha \\
& t + v_j \geq \delta_j y_j \quad \forall j \in J \\
& t \geq 0 \\
& v_j \geq 0 \quad \forall j \in J \\
& y_j \in \{0, 1\} \quad \forall j
\end{aligned} \tag{R-SIM}$$

which we refer to as R-SIM. The variables v_j and t correspond to the dual variables of (3.9) [22]. The parameter $\mathbf{\Gamma}$ controls the *price of robustness* defined as the difference between the robust optimal function value and that of SIM. The relevant range of $\mathbf{\Gamma}$ is lower bounded at zero. The maximum relevant value of $\mathbf{\Gamma}$ beyond which $\mathbf{\Gamma}$ does not affect the optimal objective value is denoted by $\hat{\mathbf{\Gamma}}$. In the derivation of such a bound, $\hat{\mathbf{\Gamma}}$ is assumed integer since in the original formulation by [22] it is interpreted as the integer number of parameters that can vary.

Proposition 3.2. *$\hat{\mathbf{\Gamma}}$ is no larger than the integer number of unimaged patient types selected by applying the greedy algorithm to SIM.*

Proof. The robust counterpart in (3.9) can be interpreted as an adversary choosing the constraint coefficients that are allowed to vary, i.e., those constraint coefficients, m_j, such that $y_j = 1$. Given the optimal solution $y_j^* \in \{0, 1\}$ to (3.9), the maximum relevant range of $\mathbf{\Gamma}$ is therefore:

$$\hat{\mathbf{\Gamma}} \leq \sum_{j=1}^{N} y_j^*, \tag{3.10}$$

showing that $\hat{\mathbf{\Gamma}}$ is bounded by the maximum number of unimaged patient types. Now consider a 0-1 knapsack problem in which the objective is to maximize the total number

of patient types not to be imaged with respect to the budget constraint as:

$$\begin{aligned} \text{maximize} \quad & \sum_{j=1}^{N} y_j \\ \text{subject to} \quad & \sum_{j=1}^{N} m_j y_j \leq \alpha \\ & y_j \in \{0,1\} \quad \forall j \end{aligned} \tag{3.11}$$

Because SIM is a restriction of (3.11), for an optimal solution $\tilde{\mathbf{y}}^*$ to (3.11), we have:

$$\hat{\mathbf{\Gamma}} \leq \sum_{j=1}^{N} y_j^* \leq \sum_{j=1}^{N} \tilde{y}_j^* \tag{3.12}$$

An upper bound on $\sum_{j=1}^{N} \tilde{y}_j^*$ can be obtained by solving the LP-relaxation of (3.11). In the LP-relaxation, we use the SIM-Greedy algorithm with a minor modification that if not imaging a patient type would cause the violation of the budget for the *split type* s defined as $s = \min\left\{j : \sum_{i=1}^{j} w_i p_i > \alpha\right\}$, we stop the algorithm and the residual budget is filled by an appropriate fractional part of type s. As the greedy choice property holds for the LP-relaxation, we have the following upper bound [86]:

$$\hat{\mathbf{\Gamma}} \leq \sum_{j=1}^{N} \tilde{y}_j^* \leq s - 1 + \frac{\left(\alpha - \sum_{j=1}^{s-1} w_j p_j\right)}{w_s p_s} \tag{3.13}$$

Because of the integrality of $\tilde{y}_j^*$, a tighter upper bound can be obtained by rounding down the solution of the greedy algorithm for the LP-relaxation:

$$\hat{\mathbf{\Gamma}} \leq \sum_{j=1}^{N} \tilde{y}_j^* \leq s - 1 \tag{3.14}$$

(3.14) shows that the maximum relevant value of $\mathbf{\Gamma}$ is at most the maximum number of unimaged patient types determined by the greedy algorithm. □

Proposition 3.2 provides a fast method to identify the relevant range of $\mathbf{\Gamma}$ in SIM. It is important to point out that the split type s is nondecreasing in the missed-rate budget α, and thus, the upper bound on $\hat{\mathbf{\Gamma}}$ is increasing in α. Moreover, when $\mathbf{\Gamma} = 0$, R-SIM is equivalent to SIM. An important implication of these findings is that the more strict the

missed-rate budget α, the more the solution is robust to uncertainty.

3.3.2. Two Imaging Tests Case

In this section, we extend our models to develop coordinated imaging protocols that consider different combinations of the imaging tests. Each branch of the decision tree in Figure 3.2 represents an imaging protocol, indexed by p, $p = 1, \ldots, 4$, and the circles represent the application of the tests, with random outcomes denoted by branches.

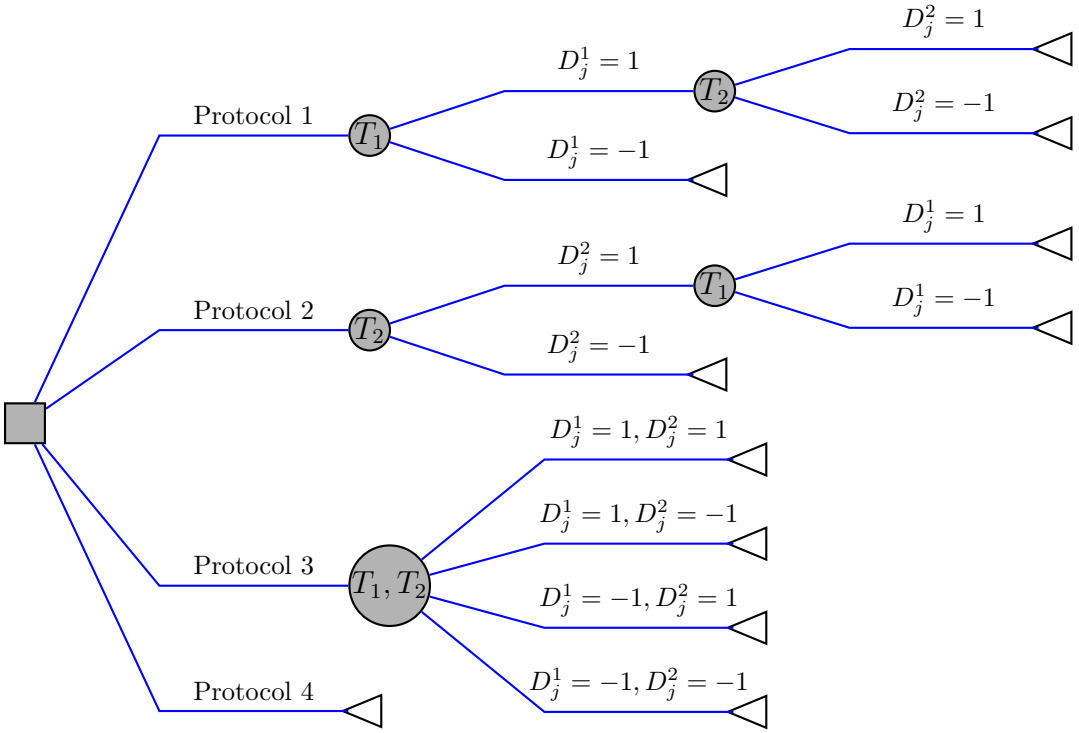

Figure 3.2: Decision tree for the design of coordinated imaging protocols for the two tests case.

We introduce a binary variable x_{jp} defined as:

$$x_{jp} = \begin{cases} 1, & \text{if patient type } j \text{ is assigned to imaging protocol } p \\ 0, & \text{otherwise} \end{cases}$$

We let n_{jp} denote the expected cost and m_{jp} denote the expected missed disease rate for patient type j under protocol p. As in the single test case, α represents the budget on the

missed disease rate for the population. The decision problem of determining the optimal assignment of patient types to imaging protocols so as to minimize the total number of imaging tests performed, can be formulated as a MCKP:

$$\min \left\{ \sum_{j=1}^{N} \sum_{p=1}^{4} n_{jp} x_{jp} \;\middle|\; \sum_{j=1}^{N} \sum_{p=1}^{4} m_{jp} x_{jp} \leq \alpha,\ \sum_{p=1}^{4} x_{jp} = 1,\ \forall j,\ \text{and } x_{jp} \in \{0,1\},\ \forall j,p \right\} \tag{3.15}$$

This model can be transformed into the standard MCKP formulation by multiplying the objective by -1, and adding the constant $\bar{n}_j = \max\{n_{jp} \mid p = 1,\ldots,4\}$ to all n_{jp} for patient type j:

$$\max \left\{ \sum_{j=1}^{N} \sum_{p=1}^{4} (\bar{n}_j - n_{jp}) x_{jp} \;\middle|\; \sum_{j=1}^{N} \sum_{p=1}^{4} m_{jp} x_{jp} \leq \alpha,\ \sum_{p=1}^{4} x_{jp} = 1,\ \forall j,\ \text{and } x_{jp} \in \{0,1\},\ \forall j,p \right\} \tag{MIM}$$

which we refer to in our context as the *multiple imaging model* (MIM).

The parameters n_{jp} and m_{jp} of MIM are determined based on the patient type-specific probabilities. We let $g_j(\cdot,\cdot)$ denote the probability of joint outcomes of imaging tests T_1 and T_2, and $g_j(\cdot \mid \cdot)$ the probability of conditional outcomes of imaging tests T_1 and T_2 for patient type j. We let c_1 and c_2 denote the costs of imaging tests T_1 and T_2, respectively. We refer to c_1 and c_2 as costs; however, they can be generalized to represent asymmetrical penalties for imaging tests on the basis of factors that differentiate imaging tests such as cost, side effects, or patient or physician preferences in different concepts. The expected cost of imaging for a patient type j type are defined as $n_{j1} = w_j(c_1 + c_2 g_j(D_j^1 = 1))$ under Protocol 1, $n_{j2} = w_j(c_2 + c_1 g_j(D_j^2 = 1))$ under Protocol 2, $n_{j3} = 2w_j(c_1 + c_2)$ under Protocol 3 and $n_{j4} = 0$ under Protocol 4. The missed disease rate are defined as $m_{j1} = w_j g_j(D_j^1 = -1, D_j^2 = 1)$ for Protocol 1, $m_{j2} = w_j g_j(D_j^1 = 1, D_j^2 = -1)$ for Protocol 2, and $m_{j4} = w_j g_j(D_j^1 = 1 \text{ or } D_j^2 = 1) = w_j(1 - g_j(D_j^1 = -1, D_j^2 = -1))$ for Protocol 4. Note that $m_{j3} = 0$ since the protocol performing both tests simultaneously has a perfect detection rate by our assumption that at least one positive test result confirms the absolute presence of metastatic disease.

Similar to the single imaging test case, we now introduce the notion of consistent risk-ordering in the context of two imaging tests. Given that our primary concern is the detection of metastatic cancer, we associate the risk of disease for a patient type with the probability of at least one positive imaging test, i.e., $g_j(D_j^1 = 1 \text{ or } D_j^2 = 1)$. Because there

are multiple imaging protocols with varying costs, we define the risk-ordering based on the expected cost of imaging protocols in the two imaging tests case.

Definition 3.2. *Given two patient types j and j' such that the risk of disease is greater for patient type j than patient type j', i.e., $g_j(D_j^1 = 1 \text{ or } D_j^2 = 1) > g_{j'}(D_{j'}^1 = 1 \text{ or } D_{j'}^2 = 1)$, the collection of decisions guiding the assignment of these patient types to imaging protocols is referred to as a consistent risk-ordering if it results in an expected cost of imaging for patient type j' that is at most as high as the expected cost of imaging for patient type j.*

We adopt the standard greedy algorithm developed for the MCKP [86]. In the greedy algorithm for MCKP, the concept of *dominance* is important in the solution of MCKP because several variables that will never be chosen in an optimal solution can be deleted a priori.

Definition 3.3. *Given two protocols s and t for patient type j, protocol s dominates t if it results in a lower cost and lower missed disease rate than protocol t. More formally:*

$$m_{js} \leq m_{jt} \quad \textit{and} \quad n_{js} \leq n_{jt} \tag{3.16}$$

Proposition 3.3. *If the results of imaging tests are conditionally independent, there is no dominance relation among the protocols for a patient type j if and only if the following condition is satisfied:*

$$(g_j(D_j^2 = -1) - g_j(D_j^1 = -1))(c_1 g_j(D_j^2 = -1) - c_2 g_j(D_j^1 = -1)) > 0 \tag{3.17}$$

Proof. Protocol 3 results in the highest cost of imaging with zero missed rate. Protocol 4 does not perform any imaging and therefore has zero cost and the highest missed rate. As illustrated in Figure 3.3, these protocols are nondominated. Therefore, we restrict our focus to Protocols 1 and 2 in the rest of this proof.

We first show that when there is no dominance, then (3.17) is satisfied. The lack of dominance implies that either $m_{j1} < m_{j2}$ and $n_{j1} > n_{j2}$, or $m_{j1} > m_{j2}$ and $n_{j1} < n_{j2}$. Consider first the case that $m_{j1} < m_{j2}$ and $n_{j1} > n_{j2}$. From the conditional independence of test results, we can express m_{j1} as:

$$\begin{aligned} m_{j1} &= w_j g_j(D_j^1 = -1)(1 - g_j(D_j^2 = -1)) \\ m_{j1} &= w_j(g_j(D_j^1 = -1) - g_j(D_j^1 = -1) g_j(D_j^2 = -1)) \end{aligned}$$

Similarly, we can express m_{j2} as $m_{j2} = w_j(g_j(D_j^2 = -1) - g_j(D_j^1 = -1)g_j(D_j^2 = -1))$. Thus, $m_{j1} < m_{j2}$ implies that $g_j(D_j^2 = -1) - g_j(D_j^1 = -1) > 0$, and $n_{j1} > n_{j2}$ implies that $c_1 g_j(D_j^2 = -1) - c_2 g_j(D_j^1 = -1) > 0$. Therefore, (3.17) is satisfied. A similar argument holds true for the case that $m_{j1} > m_{j2}$ and $n_{j1} < n_{j2}$.

We now show that when (3.17) is satisfied, then there is no dominance relation between Protocols 1 and 2. Consider first the case that $g_j(D_j^2 = -1) - g_j(D_j^1 = -1) > 0$ and $c_1 g_j(D_j^2 = -1) - c_2 g_j(D_j^1 = -1) > 0$. Based on the expressions for m_{j1} and m_{j2}, we see that $g_j(D_j^2 = -1) - g_j(D_j^1 = -1) > 0$ implies $m_{j1} < m_{j2}$. Based on the expressions for n_{j1} and n_{j2}, $c_1 g_j(D_j^2 = -1) - c_2 g_j(D_j^1 = -1) > 0$ implies that $n_{j1} > nj2$. These results show that there is no dominance relation between Protocols 1 and 2 when $g_j(D_j^2 = -1) - g_j(D_j^1 = -1) > 0$ and $c_1 g_j(D_j^2 = -1) - c_2 g_j(D_j^1 = -1) > 0$. A similar argument holds true for the case that $g_j(D_j^2 = -1) - g_j(D_j^1 = -1) < 0$ and $c_1 g_j(D_j^2 = -1) - c_2 g_j(D_j^1 = -1) < 0$, showing that $m_{j1} > m_{j2}$ and $n_{j1} < n_{j2}$. Thus, there is no dominance relation between Protocols 1 and 2. □

In a general two tests case, dominated protocols can be found according to criteria in 3.16 and can be eliminated for each patient type. We let R_j denote the set of protocols that are nondominated for type j. The size of set R_j is denoted by r_j. We assume the ordering $m_{j1} < m_{j2} < \ldots < m_{jr_j}$ in R_j.

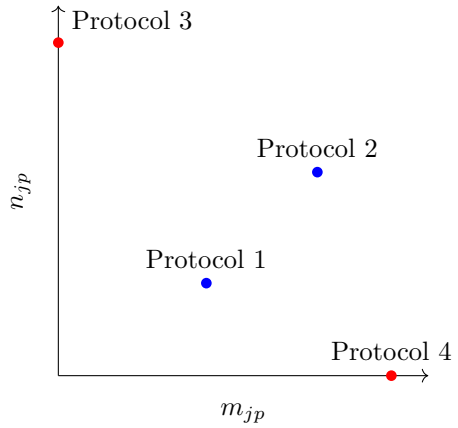

Figure 3.3: The illustration of dominance for MIM where Protocol 1 dominates Protocol 2.

In the MIM-Greedy algorithm, the incremental number of imaging tests $\tilde{n}_{jp}$ is a measure of how much we decrease imaging, and the incremental missed-rate $\tilde{m}_{jp}$ shows how much

Algorithm 5: MIM-Greedy.

For each patient type j, derive R_j and sort the protocols in R_j according to increasing missed-rate, m_{jp}. The following indices refer to protocols in R_j with respect to this order.
Construct an instance of the 0-1 knapsack problem by setting $\tilde{n}_{jp} = n_{j,p-1} - n_{jp}$ and $\tilde{m}_{jp} = m_{jp} - m_{j,p-1}$ for each R_j and $p = 2, \ldots, r_j$. Each item in this problem can be seen as 2-tuples of (j, p).
Calculate the incremental efficiencies $\tilde{e}_{jp} = \tilde{n}_{jp}/\tilde{m}_{jp}$ for each item and sort the items according to decreasing $\tilde{e}_{jp}$. With each value of $\tilde{e}_{jp}$, we associate the original indices j, p during the sorting. Use SIM-Greedy to assign patients types to protocols according to the order of sorted incremental efficiencies.
Set $x_{j1} = 1$ and $x_{jp} = 0$ for $p = 2, \ldots, r_j$ for all j.
$z^G = 2$ z^G *is the total imaging tests performed based on the current allocation*
$\overline{m} = \alpha$ $\overline{m}$ *is the residual budget*
for $\forall (j, p) \in \{\tilde{e}_{jp}\}$ **do**
    while $\tilde{m}_{jp} \leq \overline{m}$ **do**
        Assign type j *to Protocol* p;
        $\overline{m} = \overline{m} - \tilde{m}_{jp}$;
        $z^G = z^G - \tilde{n}_{jp}$;
        $x_{jp} = 1$, $x_{j,p-1} = 0$;
    end
end
return *The solution* $\mathbf{x}$ *with value* z^G.

we decrease the missed-rate budget if we assign patient type j to Protocol p instead of Protocol $p - 1$.

In addition to the MIM-Greedy algorithm, we propose the *individualized* MIM-Greedy algorithm that is consistent with the perspectives of our clinical collaborators. This algorithm assigns patient types to protocols on the basis of their estimated probability of missed metastatic disease. Each protocol is considered sequentially until one is found for which the probability of missed disease for the patient type falls below the budget α. If the probability of missed disease for a patient type under Protocol 4 is above the budget, we check the probability of missed disease for Protocols 1 and 2. If both Protocols 1 and 2 result in a probability of missed disease below the budget, we assign the type to the protocol with the lowest missed disease rate m_{jp}. It is guaranteed that the greedy solution is feasible to MIM given the properties of w_j, i.e., $w_j \in [0, 1]$ for all j and $\sum_{j=1}^{N} w_j = 1$.

Algorithm 6: Individualized MIM-Greedy.

$z^G = 0$
for $j = 1$ **to** N **do**
  if $g_j(D_j^1 = 1 \text{ or } D_j^2 = 1) \leq \alpha$ **then**
    $x_{j4} = 1$, $x_{j1} = x_{j2} = x_{j3} = 0$;
  else if $g_j(D_j^1 = -1, D_j^2 = 1) \leq \alpha$ *and* $g_j(D_j^1 = -1, D_j^2 = 1) \leq \alpha$ **then**
    if $m_{j1} \leq m_{j2}$ **then**
      $z^G = z^G + n_{j1}$, and $x_{j1} = 1$, $x_{j2} = x_{j3} = x_{j4} = 0$;
    else
      $z^G = z^G + n_{j2}$, and $x_{j2} = 1$, $x_{j1} = x_{j3} = x_{j4} = 0$;
    end
  else
    $z^G = z^G + n_{j3}$, and $x_{j3} = 1$, $x_{j1} = x_{j2} = x_{j4} = 0$;
  end
end
Return the solution $\mathbf{x}$ with value z^G.

The Robust Problem

Because the model parameters n_{jp} and m_{jp} of MIM are determined based on the probability estimates, which in turn are both affected by statistical variation, we employ the robust optimization approach discussed in Section 3.2.4. The model of data uncertainty considered for the missed-rate budget constraint of MIM is similar to SIM. Each uncertain missed-rate $\tilde{m}_{jp}$ of type j and protocol p is modeled as as independent, symmetric and bounded random variable that takes values in $[m_{jp}-\delta_{jp}, m_{jp}+\delta_{jp}]$, where δ_{jp} represents the maximum deviation ($\delta_{jp} \geq 0$) from the nominal value m_{jp}. For the uncertainty of the objective, each uncertain $\tilde{n}_{jp}$ of type j and protocol p takes values in $[\bar{n}_j - n_{jp}, \bar{n}_j - n_{jp} + \sigma_{jp}]$, where σ_{jp} ($\sigma_{jp} \geq 0$) represents the maximum deviation from the nominal value n_{jp}. Adopting the

approach in [22], the robust counterpart formulation of MIM is as follows:

$$
\begin{aligned}
&\text{maximize} \sum_{j=1}^{N} \sum_{p=1}^{4} (\bar{n}_j - n_{jp}) x_{jp} + \min_{\{S_0 \mid S_0 \subseteq J_0, |S_0| \leq \mathbf{\Gamma}_0\}} \left\{ \sum_{(j,p) \in S_0} \sigma_{jp} x_{jp} \right\} \\
&\text{subject to} \\
&\sum_{j=1}^{N} \sum_{p=1}^{4} m_{jp} x_{jp} + \max_{\substack{\{S_1 \cup \{k,l\} \mid S_1 \subseteq J_1, \\ |S_1| \leq \lfloor \mathbf{\Gamma}_1 \rfloor, \{k,l\} \in J_1 \setminus S_1\}}} \left\{ \sum_{\{j,p\} \in S_1} \delta_{jp} x_{jp} + (\mathbf{\Gamma}_1 - \lfloor \mathbf{\Gamma}_1 \rfloor) \delta_{kl} x_{kl} \right\} \leq \alpha \qquad (3.18) \\
&\sum_{p=1}^{4} x_{jp} = 1, \quad \forall j \\
&x_{jp} \in \{0,1\}, \quad \forall (j,p)
\end{aligned}
$$

where $\mathbf{\Gamma}_0$ and $\mathbf{\Gamma}_1$ are used to control the level of robustness in the objective and the missed-rate budget constraint due to the uncertainty in the model parameters, respectively. J_0 and J_1 are the sets of coefficients of the objective and missed-rate budget constraint, respectively, that are subject to uncertainty, and defined as $J_0, J_1 \subseteq \{(j,p) \mid j \in \{1,\ldots,N\} \text{ and } p \in \{1,\ldots,4\}\}$.

Remark 3.2. *Similar to the single imaging test case, the formulation proposed by Bertsimas and Sim [23] assumes that $\tilde{m}_{jp}$ is a symmetric random variable; however, in our application this is not necessarily true. This is still a reasonable assumption because the adversary seeks to maximize $\tilde{m}_{jp}$, and in practice the lower bound is not achieved. Furthermore, because both the probability of missed metastatic disease and the proportion are less than 1, the upper bound on $\tilde{m}_{jp}$ does not exceed 1 for each protocol and patient type.*

The robust counterpart in (3.18) has the following equivalent MIP formulation:

$$
\begin{aligned}
\text{maximize} \quad & \sum_{j=1}^{N}\sum_{p=1}^{4}(\bar{n}_j - n_{jp})x_{jp} + t_0\mathbf{\Gamma}_0 + \sum_{(j,p)\in J_0} u_{jp} \\
\text{subject to} \quad & \\
& \sum_{j=1}^{N}\sum_{p=1}^{4} m_{jp}x_{jp} + t_1\mathbf{\Gamma}_1 + \sum_{(j,p)\in J_1} v_{jp} \leq \alpha \\
& \sum_{p=1}^{4} x_{jp} = 1 \quad \forall j \\
& t_0 + u_{jp} \geq \sigma_{jp}x_{jp} \quad \forall (j,p) \in J_0 \\
& t_1 + v_{jp} \geq \delta_{jp}x_{jp} \quad \forall (j,p) \in J_1 \\
& t_0 \geq 0, \quad t_1 \geq 0 \\
& u_{jp} \geq 0 \quad \forall (j,p) \in J_0 \\
& v_{jp} \geq 0 \quad \forall (j,p) \in J_1 \\
& x_{jp} \in \{0,1\} \quad \forall (j,p)
\end{aligned}
\tag{R-MIM}
$$

which we refer to as R-MIM. The variables u_{jp}, v_{jp} and t_0, t_1 of R-MIM correspond to the dual variables of the linearized constraints in (3.18) [22].

3.4. Predictive Modeling

The robust optimization models described in the previous section are not limited to any one type of predictive model. Thus, we provide an example based on logistic regression (LR). LR is the most commonly used predictive modeling method in the biomedical literature. To predict the positive outcome of BS and CT scan, we utilize the binary LR models that were developed and validated in Chapter 2. In this section, we describe how we utilize the LR method to predict the probabilities of nominal imaging outcomes, and how we measure the uncertainty in predictions obtained from these models.

We develop a multinomial LR model to calculate the probabilities of joint outcomes of tests T_1 and T_2. The design matrix of the independent variables, $\mathbf{X}$, is the same as in Section 2.2.2, i.e., it contains n rows and d columns where d is the number of independent variables and the first element of each row, $\mathbf{x}_{i0} = 1$, is the intercept. We let y be the categorical dependent variable of which the categories result from the concurrent application of T_1 and T_2 tests under Protocol 3. We assume that the categories of y are coded $1, 2, 3$ or 4: category 1 corresponds to $T_1 = 1$ and $T_2 = 1$, category 2 corresponds

to $T_1 = 1$ and $T_2 = -1$, category 3 corresponds to $T_1 = -1$ and $T_2 = 1$, and category 4 corresponds to $T_1 = -1$ and $T_2 = -1$.

For the multinomial LR, we fit three independent binary LR models, in which the last outcome is chosen to be the baseline outcome and the other three outcomes are separately regressed against the baseline outcome. We estimate binary LR models, for $k < 4$, as follows:

$$\log\left(\frac{\mathbb{P}(y = k \mid \mathbf{x})}{\mathbb{P}(y = 4 \mid \mathbf{x})}\right) = \boldsymbol{\beta}_k^T \mathbf{x} \tag{3.19}$$

where $\boldsymbol{\beta}_k$ represents a set of regression coefficients for each category k with respect to the reference category 4. Exponentiating both sides of (3.19) and using the fact that all four of the probabilities must sum to one, we have:

$$\begin{aligned} \mathbb{P}(y = 4 \mid \mathbf{x}) &= \frac{1}{1 + \sum_{l=1}^{3} e^{\boldsymbol{\beta}_l^T \mathbf{x}}} \\ \mathbb{P}(y = k \mid \mathbf{x}) &= \frac{e^{\boldsymbol{\beta}_k^T \mathbf{x}}}{1 + \sum_{l=1}^{3} e^{\boldsymbol{\beta}_l^T \mathbf{x}}} \quad k < 4 \end{aligned} \tag{3.20}$$

Similar to the models in Section 2.3.2, we use LM-BFGS algorithm to find the maximum likelihood estimation (MLE) estimate of $\boldsymbol{\beta}_k$ for binary models for $k < 4$.

To measure the uncertainty of parameters in SIM and MIM, we need to measure the uncertainty in the patient type-specific probabilities. Figure 3.4 displays the distribution of individual probability estimates obtained from a LR model predicting the positive outcome of BS for patients in a (a) low-risk and (b) high-risk type. The probability estimates for the low-risk type do not diverge much from zero (negative BS), which implies that the probabilities are well-calibrated for this type. For the high-risk type, the probability estimates diverge strikingly from one (positive BS), which implies that the model can not confidently assign a prediction for this type as there are not as many patients at high-risk as at low-risk of bone metastasis.

To measure the uncertainty in predictions, we employ random sampling of the coefficient vectors of the binary and multinomial LR models based on the large-sample normal distributions of MLEs. For random sampling of coefficient vectors, we need the variances and covariances of the estimated coefficients of LR models. To illustrate the estimation of the variances and covariances, we consider a binary LR model for predicting the pos-

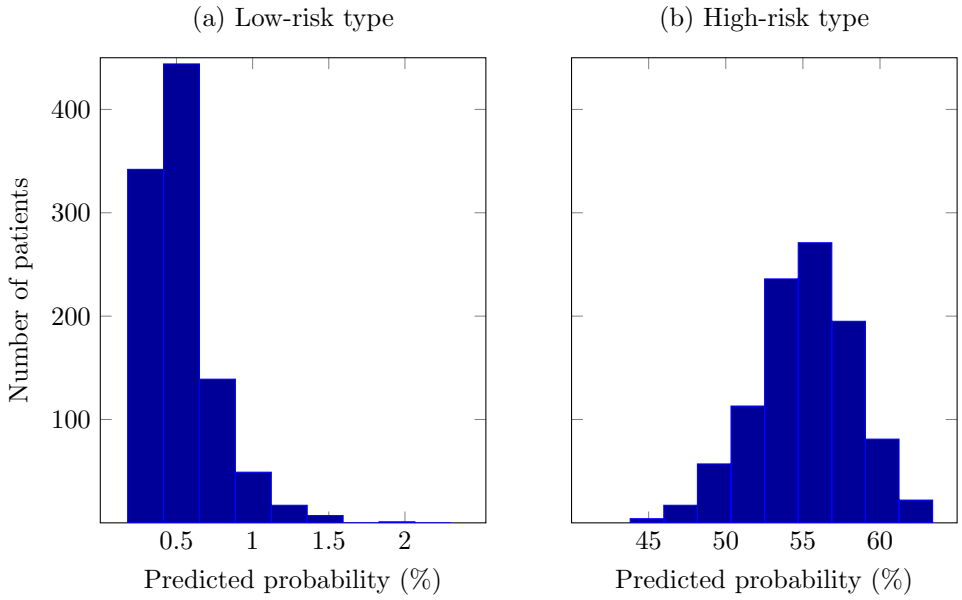

Figure 3.4: Distributions of individual probability estimates obtained from a LR model predicting the positive outcome of BS.

itive outcome of a BS: $\text{logit}\,\pi_i = \boldsymbol{\beta}^T \mathbf{x}_i$ where $\pi_i = \mathbb{P}(y_i = 1 \mid \mathbf{x}_i)$. The variances and covariances of the maximum likelihood estimates of $\boldsymbol{\beta}$ are obtained from the inverse of the so-called *observed information matrix*, denoted as $\mathbf{I}(\boldsymbol{\beta})$, i.e., $\text{Var}(\boldsymbol{\beta}) = \mathbf{I}^{-1}(\boldsymbol{\beta})$. The estimators of the variances and covariances, denoted by $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$, are obtained by evaluating $\text{Var}(\boldsymbol{\beta})$ at the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$. The information matrix can be estimated as $\hat{\mathbf{I}}(\hat{\boldsymbol{\beta}}) = \mathbf{X}^T \hat{\mathbf{V}} \mathbf{X}$, where $\mathbf{X}$ is the data matrix and $\mathbf{V}$ is a diagonal matrix defined as $\mathbf{V} = \text{diag}(\hat{\pi}_1(1-\hat{\pi}_1), \ldots, \hat{\pi}_n(1-\hat{\pi}_n))$ [78].

3.5. Results

In this section, we present results for (1) predictive modeling and (2) optimization models and greedy algorithms for the single test and multiple tests cases described in Sections 3.3.1 and 3.3.2. We used the clinical parameters that were highly associated with the positive outcome of BS and CT scan based on our study in Chapter 2 to define patient types. We chose the most commonly used categories for each of these parameters. For PSA: $\leq 4, 4-10$, $10-20$ and > 20; for GS: $< 7, = 7$ and > 7; for clinical stage: T1, T2 and T3/4 were considered. Overall, we had 36 patient types.

3.5.1. Predictive Model Results

We used the clinical datasets and variables that were described in Chapter 2.2.1 to develop and validate a multinomial LR for predicting joint outcomes of BS and CT scan. For the multinomial LR, we used a random half of the data for training and other half for validation. All models were built and evaluated with Python 2.7.11 on a HP Z230 work station with an Intel Xeon E31245W (3.4GHz) processor, 4 cores, and 16 GB of RAM. We used the `scipy.optimize` package in Python as the optimization solver.

The study population included 938 newly-diagnosed PCa patients who received both BS and CT scan at diagnosis, of which 67 (7.1%) had both tests positive, 36 (3.8%) had BS positive but CT scan negative, 40 (4.3%) had BS negative and CT scan positive, and finally, 795 (84.8%) had both tests negative. Depictions of the mean predicted value versus the true fraction of cases with $y = 1, 2$ and 3 along with the pairwise receiver operating characteristics (ROC) curves for binary models are shown in the appendix in Figure B.1. The results show good calibration in the validation samples. As expected, the binary model predicting $y = 1$ against $y = 4$ is good at discriminating patients who had both tests positive from patients who had both negative. As also seen in the calibration plots, low Brier scores indicate overall good calibration of the predicted probabilities.

To measure the statistical estimation error in the probability estimates obtained from the validated predictive models, we conducted a random sampling of $1,000$ coefficient vectors using the mean and covariances of the estimated coefficients for the binary models and multinomial model. We let $\hat{g}_j(\cdot, \cdot)$ and $\hat{g}_j(\cdot \mid \cdot)$ denote the mean patient type-specific probabilities, and let $g_j^{\min}(\cdot, \cdot)$ and $g_j^{\max}(\cdot \mid \cdot)$ denote the minimum and maximum of the patient type-specific probabilities calculated based the random samples of the coefficient vectors.

For the objective function of R-MIM, we set the nominal value of n_{j1} equal to $w_j(1 + g_j^{\min}(D_j^1 = 1))$ for Protocol 1, and n_{j2} equal to $w_j(1 + g_j^{\min}(D_j^2 = 1))$ for Protocol 2 for patient type j. Note that this corresponds to placing equal cost penalties on BS and CT scan. We set the deviation in expected number of imaging tests performed for patient type j equal to $\sigma_{j1} = w_j(g_j^{\max}(D_j^1 = 1) - g_j^{\min}(D_j^1 = 1))$ and $\sigma_{j2} = w_j(g_j^{\max}(D_j^2 = 1) - g_j^{\min}(D_j^2 = 1))$ for Protocols 1 and 2, respectively, whereas σ_{j3} and σ_{j4} are equal to zero for Protocols 3 and 4, respectively.

For the missed-rate budget constraint of both R-SIM and R-MIM, we set the nominal value of m_{j1} equal to $w_j\hat{g}_j(D_j^1 = -1, D_j^2 = 1)$ for Protocol 1, and m_{j2} to $w_j\hat{g}_j(D_j^1 = 1, D_j^2 =$

-1) for Protocol 2, and m_{j4} to $w_j\hat{g}_j(D_j^1 = 1 \text{ or } D_j^2 = 1)$ for Protocol 4. We set the deviation in missed-rate for patient type j equal to $\delta_{j1} = 0.5w_j(g_j^{\max}(D_j^1 = -1, D_j^2 = 1) - g_j^{\min}(D_j^1 = -1, D_j^2 = 1))$ for Protocol 1, and δ_{j2} to $0.5w_j(g_j^{\max}(D_j^1 = 1, D_j^2 = -1) - g_j^{\min}(D_j^1 = 1, D_j^2 = -1))$ for Protocol 2, and δ_{j4} to $0.5w_j(g_j^{\max}(D_j^1 = 1 \text{ or } D_j^2 = 1) - g_j^{\min}(D_j^1 = 1 \text{ or } D_j^2 = 1))$ for Protocol 4.

3.5.2. Optimization Model Results

In this section, we present numerical results for the single imaging test case and two imaging tests case based on the above model parameter choices. Unless otherwise specified, SIM and MIM refer to the nominal SIM and MIM. Recall that R-SIM and R-MIM refer to the robust counterpart models. To be consistent with the existing literature on knapsack problems, we proposed optimization models for both cases of the problem in maximization form in Section 3.3. For ease of interpretation, however, we present results in this section considering the minimization of average number of imaging tests to be performed rather than the true objective value. We assumed symmetrical cost penalties for BS and CT scan in our case study.

Single Imaging Test Case

Figure 3.5 illustrates that the deviations in the predictions increase with respect to the increasing probability of positive BS. We observed the same trend in the predictions for positive CT scan. These findings imply that imaging recommendations that have a consistent risk-ordering, such as solutions to the SIM-Greedy algorithm, are capable of mitigating the adverse effect of statistical error on the optimal value.

Figure 3.6 depicts the diminishing returns with respect to the increasing budget on missed disease rate for the optimal SIM and R-SIM solutions, the SIM-Greedy and individualized SIM-Greedy solutions. In this figure, R-SIM has the full protection level against statistical error, i.e., $\mathbf{\Gamma} = |J| = 36$ and thus there is no constraint on parameter variation in the uncertainty set. For both BS and CT scan, the SIM-Greedy algorithm provides solutions that are close to the optimal solutions obtained by solving SIM to optimality. At a missed-rate budget of $\alpha = 1\%$, the solutions to both SIM and SIM-Greedy reduce the average number of BSs per patient by 72.7% compared to the individualized SIM-Greedy solution. At the same missed-rate budget, the solutions to SIM and SIM-Greedy reduce the average number of CT scans per patient by 66.7% and 60.6%, respectively, compared

to the individualized SIM-Greedy solution. At a missed-rate budget of 2%, the solutions to SIM and SIM-Greedy reduce the average number of BSs per patient by 81.2% and 77.9%, respectively, compared to the individualized SIM-Greedy solution. At the same missed-rate budget, the solutions to SIM and SIM-Greedy reduce the average number of CT scans per patient by 79.5% and 75.7%, respectively, compared to the individualized SIM-Greedy solution. Thus, the results suggest that the SIM-Greedy algorithm provides near optimal solutions while the individualized SIM-Greedy algorithm has poor performance.

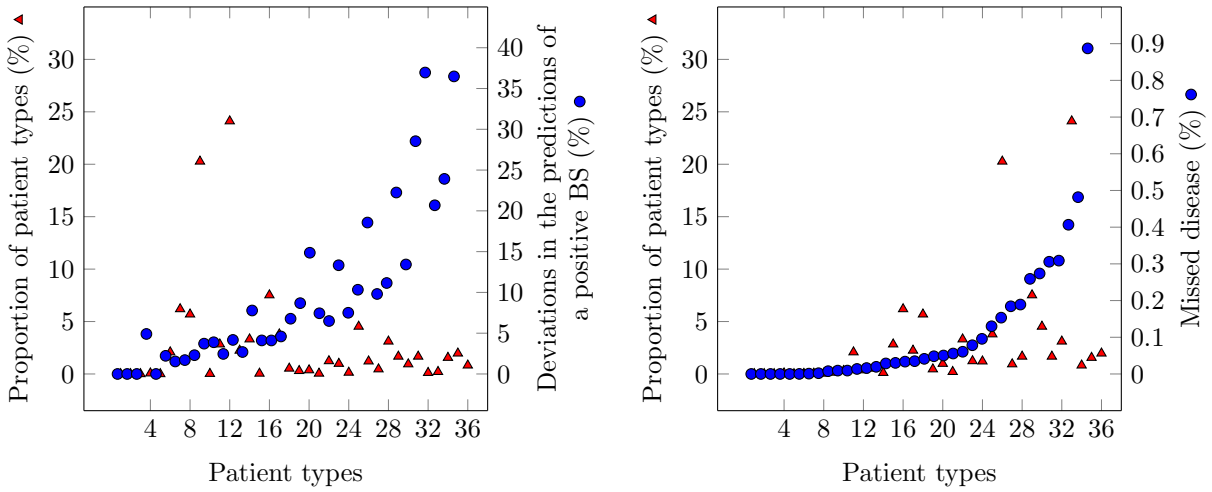

Figure 3.5: Left: Illustration of the relation between the proportion of patient types and the variations in the estimated probability of positive BS. The patient types are sorted in the order of increasing risk of disease. Right: Illustration of the relation between the proportion of patient types and the expected rate of missed disease. The patient types are sorted in the order of increasing rate of missed disease.

In Chapter 2, we developed a state-wide imaging criteria that has been implemented by MUSIC. This imaging criteria was found to be Pareto optimal with respect to the expected number of positive outcomes missed and expected number of negative tests based on the patients who received an imaging test at diagnosis. To evaluate the performance of the MUSIC criteria on the basis of expected missed disease rate and expected number of imaging test performed per patient at the population level, we assigned patient types to receive or not receive a BS or CT scan according to the MUSIC criteria. Figure 3.6 demonstrates that the MUSIC criteria is on the efficient frontier for both BS and CT scan.

To examine the effect of the protection level, $\mathbf{\Gamma}$, on the tradeoff between robustness and optimality, we consider the probability of missed-rate budget violation by randomly

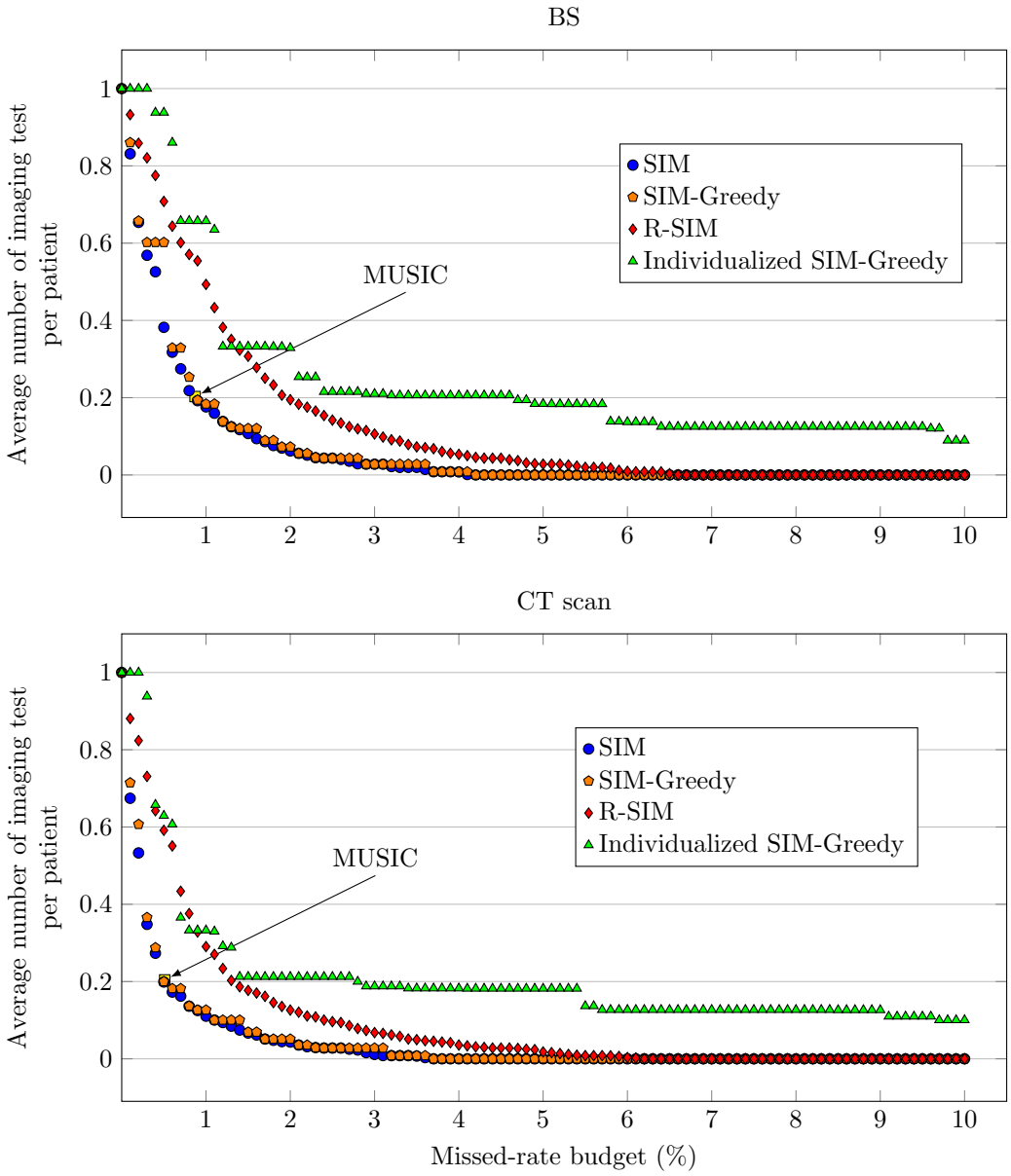

Figure 3.6: The diminishing returns from the optimal SIM and R-SIM solutions, the SIM-Greedy and individualized SIM-Greedy solutions as a function of the increasing missed-rate budget for BS and CT scan. R-SIM represents the optimal value with full protection against statistical variation.

sampling coefficient vectors of the binary LR models, as explained in Section 3.4. At a certain missed-rate budget, we construct instances of SIM using $1,000$ randomly selected coefficient vectors of the binary LR models. We estimate the probability of missed-rate

budget violation as the fraction of SIM instances that are infeasible by the original optimal solutions of R-SIM with various protection levels. To examine the effect of sampling variation on the model outcomes, we obtain 30 independent samples of 1000 coefficient vectors of the binary LR models and constructed 95% confidence interval (CI)s.

Figure 3.7 illustrates the effect of the protection level $\mathbf{\Gamma}$ on the objective function value and the probability of missed-rate budget violation for R-SIM. We observe that as the protection level increases, the probability of missed-rate budget violation decreases while the number of imaging test per patient increases. Based on Proposition 3.2, we found the upper bound on maximum relevant value $\hat{\mathbf{\Gamma}}$ by applying the SIM-Greedy algorithm on SIM for BS and CT scan as 23 and 25, respectively, at a missed-rate budget of 1%. Because we solve SIM and R-SIM as minimization problems for easier interpretation of the results, these upper bounds on $\hat{\mathbf{\Gamma}}$ correspond to 13 and 11 for BS and CT scan, respectively. Figure 3.7 confirms that $\hat{\mathbf{\Gamma}} < 10$ for both BS and CT scan demonstrating that the true bounds for these particular instances are tighter than the bound in Proposition 3.2. In the absence of protection to the missed-rate budget constraint, the optimal value is 0.18 and 0.11 for BS and CT scan, respectively. However, with maximum protection, the number of imaging tests per patient is increased to 0.49 and 0.29 for BS and CT scan, respectively, indicating a more conservative solution that recommends more imaging when $\mathbf{\Gamma}$ is unbounded.

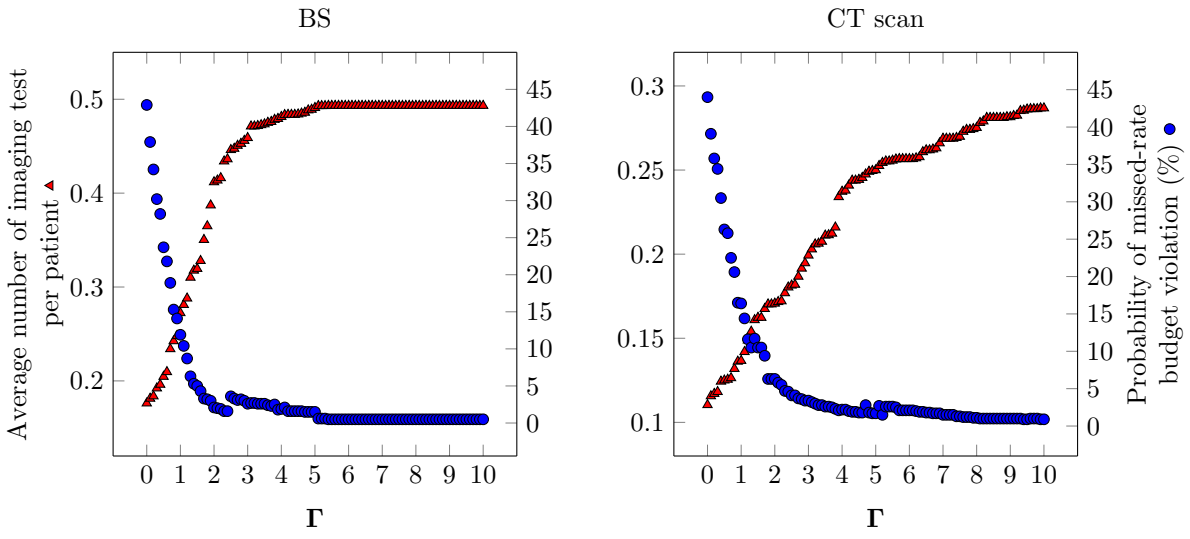

Figure 3.7: The tradeoff between the optimal value and the probability of missed-rate budget violation of R-SIM as a function of the increasing protection level $\mathbf{\Gamma}$ at a budget of 1% for BS and CT scan.

In Table 3.1, we present a sample of the objective function value and the probability of missed-rate budget violation for R-SIM at budgets of 1% and 2%. At a 1% missed-rate budget, the solution to R-SIM with $\mathbf{\Gamma} = 0.5$ and 1 decrease the probability of missed-rate budget violation by 48.2% and 79.6%, respectively, while increasing the average number of BSs per patient by 20.5% and 64.8%, respectively. At the same missed-rate budget, the solutions to R-SIM with $\mathbf{\Gamma} = 0.5$ and 1 decrease the probability of missed-rate budget violation by 47.5% and 73.1%, respectively, while increasing the average number of CT scans per patient by 14.3% and 43.8%, respectively. At a 2% missed-rate budget, the solutions to R-SIM with $\mathbf{\Gamma} = 0.5$ and 1 decrease the probability of missed-rate budget violation by 40.1% and 67.8%, respectively, while increasing the average number of BSs per patient by 17.5% and 42.9%, respectively. At the same missed-rate budget, the solutions to R-SIM with $\mathbf{\Gamma} = 0.5$ and 1 decrease the probability of missed-rate budget violation by 24.3% and 56.7%, respectively, while increasing the average number of CT scans per patient by 7.0% and 23.3%, respectively. These results suggest that the robust formulation offers flexibility to the decision maker when trading off between the competing goals of minimizing the number of imaging tests performed in the population and minimizing the probability of missed-rate budget violation.

In Section 3.3.1, it was suggested that the SIM-Greedy algorithm tends to provide a robust solution by selecting patient types for no imaging from high to low risk of disease. As shown in Table 3.1, the SIM-Greedy solution provides more protection against missed-rate budget violation than the optimal SIM solution for both BS and CT scan. At a 1% missed-rate budget, the SIM-Greedy solution reduces the probability of missed-rate budget violation by 17.1% while increasing the average number of BSs per patient by 3.8% compared to the optimal SIM solution. At a 2% missed-rate budget, the SIM-Greedy solution reduces the probability of missed-rate budget violation by 40.3% while increasing the average number of BSs per patient by 15.9%. At a 1% missed-rate budget, the SIM-Greedy solution reduces the probability of missed-rate budget violation by 39.4% while resulting in an increase of 10.6% in the average number of CT scans per patient. At a 2% missed-rate budget, the reduction in the probability of missed-rate budget violation is 58.0% with an increase of 15.7% in the average number of CT scans per patient compared to the optimal SIM solution. These results suggest that the SIM-Greedy algorithm generates more robust solutions as anticipated.

The SIM-Greedy algorithm provides greater reduction in the probability of missed-rate budget violation for CT scan than BS compared to SIM. As illustrated in Figure B.4 on a

random sample of coefficient vectors, patients tend to have higher probability of a positive CT scan than a positive BS. For the five patient types with highest risk of disease, the mean probability of positive CT scan is 40.8% (ranging from 25.3% to 70.9%), whereas the mean probability of positive BS is 34.0% (ranging from 17.7% to 52.0%). Moreover, these patient types are subject to much higher statistical variation than patient types at low risk of disease (illustrated in Figure 3.5). As the SIM-Greedy algorithm selects these patient types for no imaging at missed-rate budgets of 1% and 2%, it provides more protection against missed-rate budget violation than the optimal SIM solutions for CT scan than BS.

Compared to the MUSIC imaging criteria, the SIM-Greedy results in a lower average number of BS and CT scan per patient at missed-rate budgets of 1% and 2%, and the reduction in the average number of imaging tests per patient by the SIM-Greedy algorithm becomes more significant as the missed-rate budget increases; however, the MUSIC criteria solution provides more protection against missed-rate budget violation at each budget level. For example, at a 2% missed-rate budget, the SIM-Greedy solutions reduce the average number of BSs and CT scans per patient by 64.0% and 75.0%, respectively, compared to the MUSIC criteria solution; however, they result in significantly higher probability of missed-rate budget violation (26.2% and 14.6%, respectively) than the MUSIC criteria solutions for BS and CT scan (0.4% and 0.1%, respectively).

The solution to the MUSIC imaging criteria results in 0.20 BSs per patient with 26.6% probability of missed-rate budget violation at a budget of 1%. At the same missed-rate budget, the solution to R-SIM with $\mathbf{\Gamma} = 1$ reduces the probability of missed-rate budget violation by 67.1% to 8.7% while increasing the number of BSs per patient by 42.9% to 0.29 compared to the MUSIC solution. At a 2% missed-rate budget, the solution to R-SIM with $\mathbf{\Gamma} = 2$ reduces the number of BSs per patient by 42.4% to 0.12; however, increases the probability of missed-rate budget violation from 0.4% to 5.5%. For CT scan, the MUSIC imaging criteria provides a good solution at a missed-rate budget of 1%, i.e., there is no R-SIM resulting in a lower probability of missed-rate budget violation and a lower number of CT scans per patient than the MUSIC criteria. At a 2% missed-rate budget, the solution to R-SIM with $\mathbf{\Gamma} = 2$ reduces the number of CT scans per patient by 65.7% from 0.20 to 0.07 while increasing the probability of missed-rate budget violation from 0.08% to 5.4%, compared to the MUSIC criteria solution. At the same missed-rate budget, the solution to R-SIM with $\mathbf{\Gamma} = 6$ reduces the number of CT scans per patient by 40.7% to 0.12 while increasing the probability of missed-rate budget violation to 0.6%. These findings show that depending on the risk attitude of the decision maker, R-SIM with various choices of

$\mathbf{\Gamma}$ can reduce the number of CT scans per patient substantially compared to the MUSIC criteria.

Table 3.1: Comparison of imaging solutions for BS and CT scan at missed-rate budgets of 1% and 2%.

	Probability of missed-rate budget violation (%)	Optimal value	Probability of missed-rate budget violation (%)	Optimal value
BS				
	$\alpha = 1\%$		$\alpha = 2\%$	
MUSIC	26.60 (25.98 − 27.22)	0.203	0.40 (0.33 − 0.46)	0.203
SIM-Greedy	35.55 (34.94 − 36.16)	0.184	26.22 (25.65 − 26.79)	0.073
Individualized SIM-Greedy	0.01 (0.0 − 0.02)	0.657	0.02 (0.0 − 0.03)	0.329
R-SIM				
$\mathbf{\Gamma} = 0.0$	42.88 (42.38 − 43.38)	0.176 (0.177 − 0.178)	43.89 (43.35 − 44.44)	0.063 (0.063 − 0.064)
$\mathbf{\Gamma} = 0.2$	33.94 (33.31 − 34.56)	0.187 (0.186 − 0.189)	36.23 (35.49 − 36.97)	0.068 (0.068 − 0.069)
$\mathbf{\Gamma} = 0.5$	22.20 (21.17 − 23.23)	0.212 (0.207 − 0.216)	26.27 (25.53 − 27.01)	0.074 (0.072 − 0.075)
$\mathbf{\Gamma} = 0.7$	15.74 (14.72 − 16.76)	0.242 (0.235 − 0.248)	21.06 (20.06 − 22.06)	0.081 (0.079 − 0.083)
$\mathbf{\Gamma} = 1.0$	8.74 (7.82 − 9.66)	0.290 (0.280 − 0.299)	14.13 (13.37 − 14.88)	0.090 (0.088 − 0.091)
$\mathbf{\Gamma} = 2.0$	2.67 (2.24 − 3.10)	0.438 (0.425 − 0.451)	5.48 (4.85 − 6.11)	0.117 (0.113 − 0.121)
$\mathbf{\Gamma} = 6.0$	0.69 (0.58 − 0.79)	0.514 (0.504 − 0.524)	0.87 (0.74 − 1.01)	0.179 (0.171 − 0.188)
CT scan				
MUSIC	3.23 (2.97 − 3.49)	0.204	0.08 (0.04 − 0.12)	0.204
SIM-Greedy	25.37 (23.04 − 27.49)	0.125	14.57 (14.14 − 15.00)	0.051
Individualized SIM-Greedy	0.68 (0.59 − 0.78)	0.332	0.06 (0.03 − 0.09)	0.212
R-SIM				
$\mathbf{\Gamma} = 0.0$	41.89 (41.36 − 42.41)	0.112 (0.113 − 0.114)	34.72 (34.15 − 35.30)	0.043 (0.043 − 0.044)
$\mathbf{\Gamma} = 0.2$	32.76 (31.70 − 33.82)	0.119 (0.120 − 0.121)	34.18 (33.51 − 34.84)	0.043 (0.043 − 0.044)
$\mathbf{\Gamma} = 0.5$	21.98 (20.67 − 23.29)	0.128 (0.131 − 0.134)	26.29 (24.96 − 27.62)	0.046 (0.045 − 0.047)
$\mathbf{\Gamma} = 0.7$	16.85 (15.38 − 18.33)	0.137 (0.141 − 0.146)	21.13 (19.76 − 22.50)	0.048 (0.047 − 0.050)
$\mathbf{\Gamma} = 1.0$	11.28 (10.06 − 12.51)	0.161 (0.157 − 0.166)	15.02 (13.50 − 16.55)	0.053 (0.051 − 0.055)
$\mathbf{\Gamma} = 2.0$	3.68 (3.03 − 4.33)	0.230 (0.199 − 0.261)	5.38 (4.59 − 6.16)	0.070 (0.066 − 0.074)
$\mathbf{\Gamma} = 6.0$	0.73 (0.57 − 0.89)	0.360 (0.328 − 0.392)	0.61 (0.50 − 0.72)	0.121 (0.113 − 0.129)

The numbers in the parentheses represent the 95% CIs calculated based on the 30 independent samples of 1000 coefficient vectors of the LR models.

As illustrated in Table 3.1, there is a clear trade-off between the level of protection against the statistical estimation error and the value of the objective function. To investigate how this tradeoff is affected when the protection level is changed at varying missed-rate budgets, we consider a measure for the cost of robustness, referred to as the *price of robustness*. It is defined as the difference between the worst-case objective function value of the robust solution and the objective function value of the nominal solution [23]. For instance, assuming that z^{SIM} and $z^{\text{R-SIM}}$ are the objective values of SIM and R-SIM, respectively, the price of robustness is determined as follows:

$$\text{Price of robustness} = \frac{z^{\text{R-SIM}} - z^{\text{SIM}}}{z^{\text{R-SIM}}} \times 100\%$$

Figure 3.8 demonstrates that the average price of robustness increases with respect to the increasing protection level for R-SIM at missed-rate budgets of 1%, 2% and 3%, based on the random samples of coefficient vectors. For BS, the price of robustness at missed-rate budgets of 2% and 3% are slightly lower than of a 1% budget for $\mathbf{\Gamma} < 8$. For $\mathbf{\Gamma} > 8$, the maximum increase in the price of robustness between missed-rate budgets of 1% and 2% is 46.5% whereas it is 78.8% between budgets of 2% and 3%. For CT scan, the price of robustness is lower at a 2% missed-rate budget than a 1% budget for $\mathbf{\Gamma} < 10$, and the maximum increase in the price of robustness for $\mathbf{\Gamma} > 10$ is 32.8%. However, the price of robustness is always higher at a 3% missed-rate budget than 1% and 2% budgets. The increase in the price of robustness between missed-rate budgets of 2% and 3% can be as high as 351.2%. These findings suggest that the consideration of an uncertainty set is crucial in the assessment of diagnostic testing decisions when the model parameters are affected by statistical estimation error inherent in predictions.

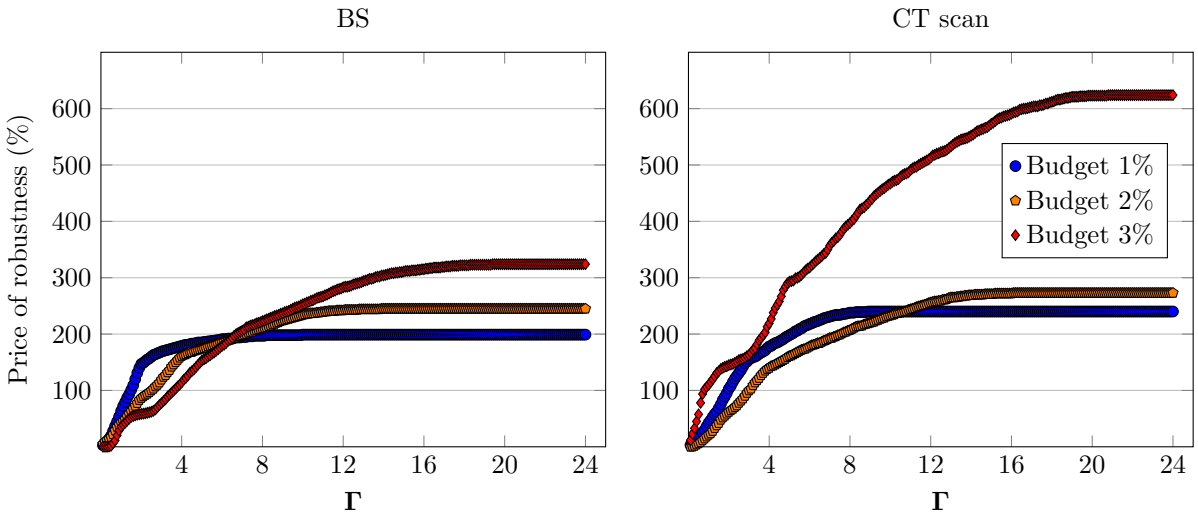

Figure 3.8: The price of robustness at varying missed-rate budgets for BS and CT scan.

Two Imaging Tests Case

Figure 3.9 depicts the diminishing returns with respect to the increasing budget on missed disease rate for the optimal MIM and R-MIM solutions, the MIM-Greedy and individualized MIM-Greedy solutions. In this figure, R-MIM has the full protection level against statistical variation, i.e., $\mathbf{\Gamma}_0 = |J_0| = 72$ and $\mathbf{\Gamma}_1 = |J_1| = 108$. Both the optimal MIM

solutions and the MIM-Greedy solutions reduce the average number of imaging tests per patient significantly compared to the individualized MIM-Greedy solutions. At a missed-rate budget of 1%, both the optimal MIM solution and the MIM-Greedy solution reduce the average number of imaging tests per patient by 55.7% compared to the individualized MIM-Greedy solution. At a 2% missed-rate budget, the optimal MIM solution and the MIM-Greedy solution reduce the average number of imaging tests per patient by 58.4% and 55.4%, respectively, compared to the individualized MIM-Greedy solution. These results suggest that the individualized MIM-Greedy algorithm performs very poorly on the basis of mean imaging per patient for a given missed-rate budget. However, the individualized MIM-Greedy algorithm guarantees all patient types have an actual missed-rate that falls below the missed-rate budget α. Thus, the difference between the optimal MIM solution and the individualized MIM-Greedy solution can be viewed as the population benefit from "central planning".

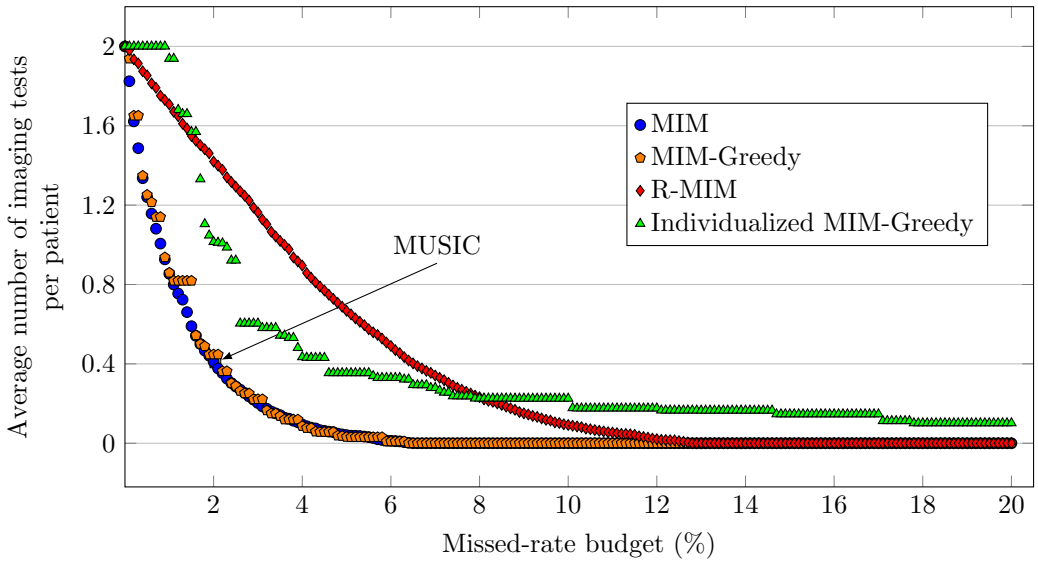

Figure 3.9: The optimal values of MIM, R-MIM, MIM-Greedy and individualized MIM-Greedy algorithms as a function of the increasing missed-rate budget. R-MIM represents the optimal value with full protection against statistical variation.

To evaluate the robustness of coordinated imaging protocols, we conduct sensitivity analyses by varying the values of the protection levels $\mathbf{\Gamma}_0$ and $\mathbf{\Gamma}_1$ in the sets $\{0, 4, 8, \ldots, 72\}$ and $\{0, 4, 8, \ldots, 108\}$, respectively. Figures B.2 and B.3 show that the optimality and robustness are not affected by the changes in $\mathbf{\Gamma}_0$ values, which controls the level of conservatism in the objective function of R-MIM. Because the patient types at high risk of disease con-

stitute a small portion of the population but are associated with high deviations in the probability estimates for positive imaging tests (illustrated in Figure 3.5), it neutralizes the impact of the protection level $\mathbf{\Gamma}_0$ on the robustness of the optimal R-MIM solution. Figure 3.10 demonstrates that the tradeoff between the robustness and optimality of coordinated imaging largely depends on the protection level $\mathbf{\Gamma}_1$, and the maximum relevant value $\hat{\mathbf{\Gamma}}_1 < 8$ at missed-rate budgets of 1% and 2%. Moreover, there are choices of $\mathbf{\Gamma}_1$ for which the optimal R-MIM solution provides substantial protection against missed-rate budget violation without heavily penalizing the optimal value.

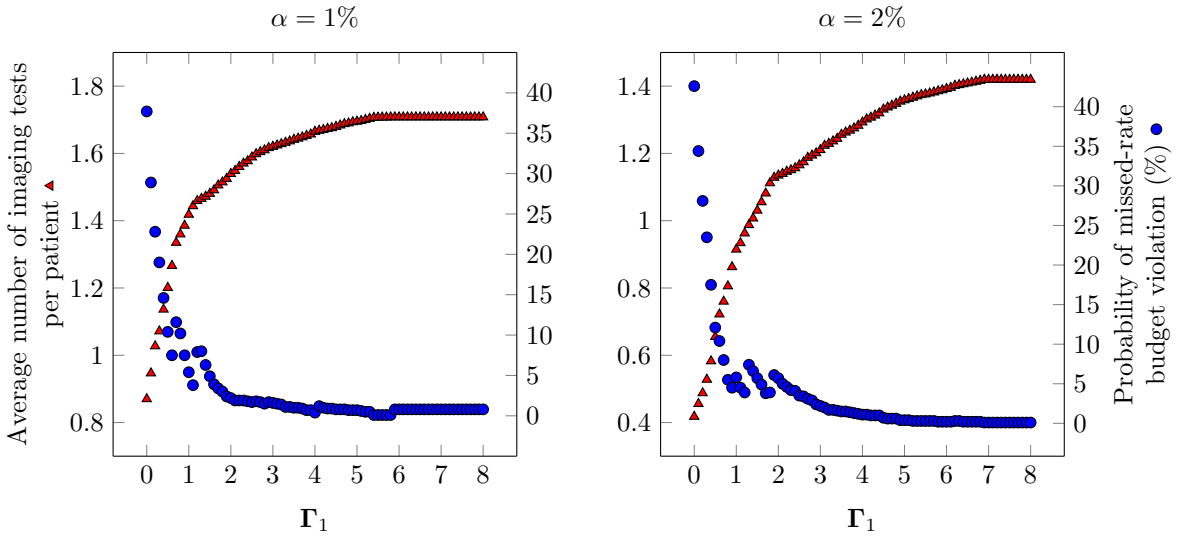

Figure 3.10: The tradeoff between the optimal value and the probability of missed-rate budget violation of R-MIM as a function of the increasing protection level $\mathbf{\Gamma}_1$.

To better illustrate the tradeoff between efficient imaging and protection against missed-rate budget violation, Table 3.2 presents a sample of the objective function value and the probability of missed-rate budget violation for R-MIM at budgets of 1% and 2%. At a missed-rate budget of 1%, the optimal MIM solution and the MIM-Greedy solution exhibit similar performance in terms of the probability of missed-rate budget violation and the optimal value. At a 2% missed-rate budget, the MIM-Greedy solution, however, reduces the probability of missed-rate budget violation by 20.3% while increasing the average number of imaging tests per patient by 8.7% compared to the optimal MIM solution.

The solution to the MUSIC criteria results in a very high probability of missed-rate budget violation at a missed-rate budget of 1% as the solution is estimated to miss 2% of

Table 3.2: Comparison of imaging solutions at missed-rate budgets of 1% and 2%.

	Probability of missed-rate budget violation (%)	Optimal value	Probability of missed-rate budget violation (%)	Optimal value
	$\alpha = 1\%$		$\alpha = 2\%$	
MUSIC	98.32 (98.19 – 98.45)	0.407	45.45 (44.80 – 46.10)	0.407
MIM-Greedy	38.49 (37.05 – 39.92)	0.848 (0.845 – 0.851)	33.43 (32.28 – 34.56)	0.448 (0.447 – 0.449)
Individualized MIM-Greedy	0.0	1.938	1.85 (1.70 – 2.00)	1.017 (1.014 – 1.020)
R-MIM				
$\mathbf{\Gamma}_1 = 0.0$	38.62 (38.22 – 39.03)	0.861 (0.847 – 0.864)	41.95 (41.54 – 42.36)	0.409 (0.406 – 0.413)
$\mathbf{\Gamma}_1 = 0.2$	24.52 (23.16 – 25.88)	0.982 (0.967 – 0.997)	31.28 (30.28 – 32.27)	0.467 (0.461 – 0.472)
$\mathbf{\Gamma}_1 = 0.5$	14.52 (13.23 – 15.80)	1.142 (1.115 – 1.169)	18.39 (16.69 – 20.09)	0.571 (0.547 – 0.595)
$\mathbf{\Gamma}_1 = 0.7$	9.98 (8.75 – 11.20)	1.237 (1.208 – 1.266)	13.67 (12.17 – 15.16)	0.662 (0.630 – 0.694)
$\mathbf{\Gamma}_1 = 1.0$	9.60 (7.84 – 11.35)	1.368 (1.343 – 1.393)	8.74 (7.66 – 9.83)	0.782 (0.752 – 0.811)
$\mathbf{\Gamma}_1 = 2.0$	5.54 (4.77 – 6.31)	1.489 (1.474 – 1.505)	3.12 (2.80 – 3.44)	1.038 (1.008 – 1.068)
$\mathbf{\Gamma}_1 = 6.0$	0.51 (0.41 – 0.61)	1.655 (1.640 – 1.670)	0.58 (0.47 – 0.69)	1.288 (1.256 – 1.319)

The numbers in the parentheses represent the 95% CIs calculated based on the 30 independent samples of 1000 coefficient vectors of the LR models. The protection level $\mathbf{\Gamma}_0$ is set to its maximum (i.e., $|J_0| = 72$) in R-MIM.

the metastatic cases in the population. At a missed-rate budget of 2%, the MIM-Greedy solution provides a 26.4% reduction in the probability of missed-rate budget violation with a 10.0% increase in the average number of imaging tests per patient compared to the MUSIC criteria solution. At the same missed-rate budget, the optimal solutions to R-MIM with $\mathbf{\Gamma}_1 = 0.2$ and $\mathbf{\Gamma}_1 = 0.5$ provide 31.2% and 59.5% reductions in the probability of missed-rate budget violation, respectively, while increasing the average number of imaging tests per patient by 14.7% and 40.3%, respectively. These findings suggest that the optimal MIM solutions and the MIM-Greedy solutions offer the decision maker the capability to adjust the tradeoff between the degree of conservatism of the solution and reduction in the total number of imaging tests performed.

Similar to the single test case, we now investigate how the price of robustness is affected by the changes in the protection level $\mathbf{\Gamma}_1$ at varying missed-rate budgets. Figure 3.11 illustrates that the price of robustness is similar for small values of $\mathbf{\Gamma}_1$ at missed-rate budgets of $1\%, 2\%$ and 3%. Moreover, both the maximum cost of robustness and the maximum relevant value of $\hat{\mathbf{\Gamma}}_1$ increase as the budget on missed rate increases. The high cost of robustness is driven by the high variations that exist in the expected missed disease rates. As illustrated in Figure B.5, some protocols have very high deviations compared to the nominal missed rates, e.g., the deviation from a nominal missed rate can be 9 times more than the nominal rate for some patient types.

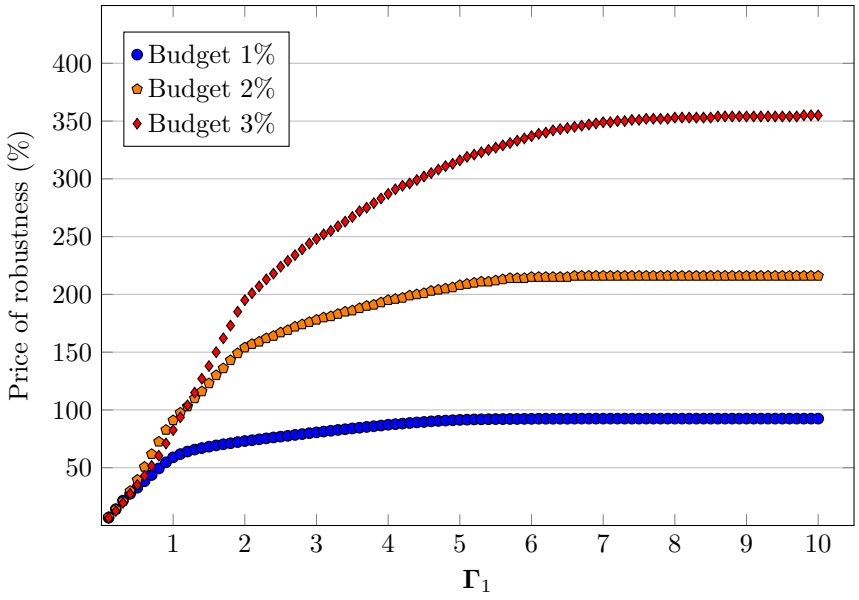

Figure 3.11: The price of robustness at varying missed-rate budgets.

Independent versus Coordinated Imaging

In this section, we investigate benefits of coordinated imaging over optimizing imaging independently for BS and CT scan. To do this, we solve SIM for both BS and CT scan, and use the optimal solutions from these models to create a solution to MIM. We apply the following rule in generating a solution to MIM: if type j is assigned to both BS and CT scan, then $x_{j3} = 1$; if type j is assigned to BS but not to CT scan, then $x_{j1} = 1$; if type j is assigned to CT scan but not to BS, then $x_{j2} = 1$; else $x_{j4} = 1$ for each patient type j, $j = 1, \ldots, N$. Next, we determine the expected number of imaging tests performed and the resulting missed disease rate in the population using the solution we create to MIM.

Figure 3.12 shows that the optimal SIM solutions for BS and CT scan, when evaluated in MIM, result in a lower number of imaging tests performed than the optimal MIM solution at varying budgets on missed rate. However, Figure 3.12 also shows that the optimal SIM solutions, when evaluated in MIM, result in a higher missed disease rate in the population than the missed-rate budget, therefore, yielding infeasible solutions to MIM.

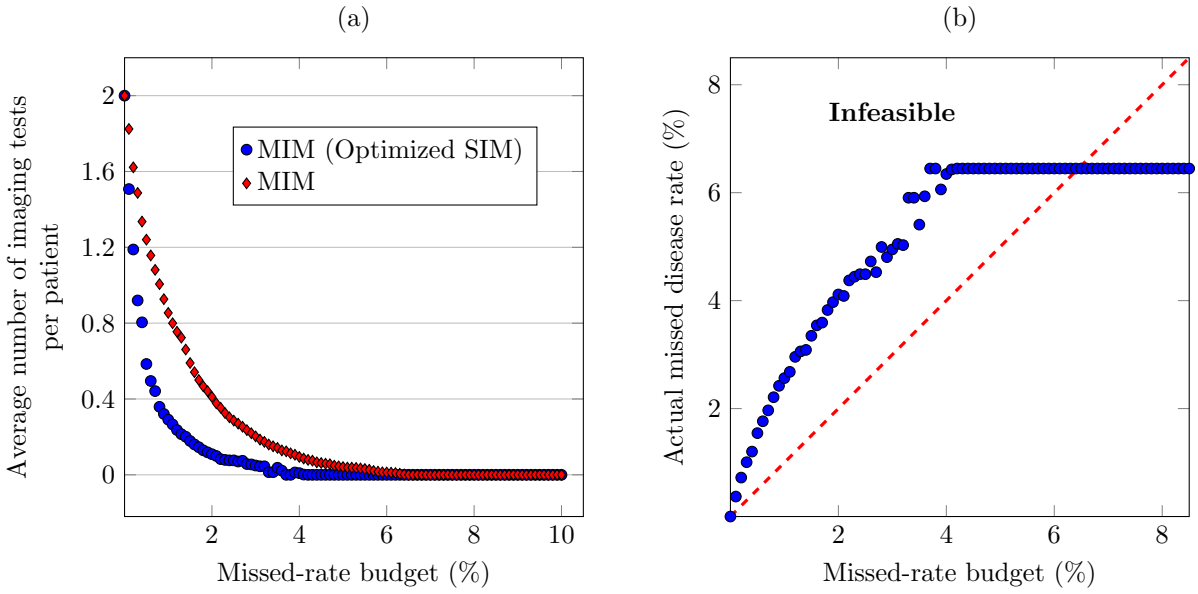

Figure 3.12: The optimal SIM solutions for BS and CT scan results in a lower number of imaging tests performed but higher missed disease rate than the budget, when evaluated in MIM.

3.6. Conclusions

The decision of how to utilize imaging tests (BS and CT scan) for PCa staging effectively and efficiently is important because (1) failure to detect metastatic disease leads to serious health outcomes including increased morbidity and mortality, and (2) over-imaging leads to more tests and treatments that are unlikely to benefit the patient and in the worst case can cause harm from false positives and radiation exposure, and increase in healthcare costs. Motivated by the lack of a holistic clinical perspective that integrates imaging decisions for PCa staging, in this chapter we studied the optimal design of robust coordinated imaging protocols under budget restrictions on the percentage of patients with missed disease in the population. We modeled this problem within an optimization framework that incorporates the perspectives of patients and physicians at the population level, and we proposed models for sequential testing where the outcome of one test informs the decision about the follow-up test. We integrated predictive modeling and robust optimization methods to account for errors in predictions. In addition to the optimization-based models, we proposed clinically motivated heuristics and evaluated the worst-case and average case behavior of these heuristics.

We studied two cases of the problem: single imaging test case and two imaging tests case. In both cases of the problem, the models we developed aim to reduce imaging burden while limiting the percentage of the population with missed disease to a certain missed-rate budget predefined by the decision maker. Our case study on real medical data suggested that both the robust optimization models and greedy algorithms offer the decision maker the capability to reduce the risk of missed-rate budget violation significantly without greatly comprising the optimal value.

In the single imaging test case, the greedy solution offered substantial safety benefits against the missed-rate budget violation compared to the deterministic optimal solution for both BS and CT scan. Moreover, the risk reduction became more significant when the probability estimates for a positive test were subject to high statistical estimation error like in the case of CT scan. These findings suggested that imaging recommendations that have a consistent risk-ordering can be developed by implementing a simple greedy algorithm in the presence of a single test. We also found that although not as significant as in the single test case, the greedy algorithm resulted in lower risk of missed-rate budget violation than the deterministic model in the two imaging tests case.

For both cases of the problem, the price of robustness was high because of high statistical estimation error, rendering the incorporation of robust optimization models into clinical decision making useful to trade off the protection against missed-rate budget violation with the number of imaging tests performed in the population. Moreover, we showed that the coordinated imaging in PCa staging is more beneficial than the optimized single imaging. The coordinated imaging offers the potential to achieve better health outcomes in the population while reducing imaging tests performed. Hence, these models are particularly relevant for clinical decision making with implications for patients and physicians. Our results also shows that optimizing at the population level (i.e., central planning) may differ significantly from optimizing for each patient type independently. This raises important questions abut how to trade off between the different perspectives of patients and physicians.

Chapter 4.

Decision Analysis for Assessment of Long Term Outcomes Associated with Newly Discovered Biomarkers at Repeat Biopsy

4.1. Introduction

In this chapter, we shift the focus from imaging tests used to diagnose the spread of cancer to continuous biomarker tests for early detection of prostate cancer (PCa) in men that have not yet been diagnosed. This is an important public health concern because of the propensity for men to develop PCa, and the dangers of PCa going undetected. It is also an excellent case study because of the large number of new biomarker tests that have been developed in recent years for screening and diagnosis of PCa. Since no single biomarker on its own is considered satisfactory, one approach to improve early diagnosis of PCa is to combine the information of several biomarker tests together. This chapter is based on the findings reported in our paper [109].

In recent years, new blood serum, urine, and radiologic biomarkers have been discovered that have the potential to improve early diagnosis of PCa [119]. Biomarker tests are distinguished clinically based on their sensitivity and specificity. The former is the probability a test is positive given the disease is present, and the latter is the probability the test is negative given the disease is not present. In dealing with a continuous biomarker, it is

necessary to define thresholds for determining when to employ the additional biomarkers and when to carry out the gold standard test (biopsy). Setting the threshold to a low value results in high sensitivity but also false positive results that cause anxiety and unnecessary referral of patients for costly and invasive procedures, such as biopsies. Setting the threshold to a high value, on the other hand, leads to false negative results that cause a disease to go undetected, and potentially progress to a life threatening stage.

Commonly used diagnostic indicators for the early detection of PCa include abnormal digital rectal examination (DRE) and an elevated prostate-specific antigen (PSA). Serum PSA levels above 2.5 to 4 ng/ml and/or suspicious DRE may indicate the presence of PCa; however, the performance of PSA alone with a cutoff of 4 ng/ml is reported to yield a positive predictive value of only $24-37\%$ [33, 160], and up to 75% of these men have a negative first biopsy. Furthermore, PCa is detected in $10-35\%$ of men with negative first biopsy [132, 145]. In clinical practice, it is often uncertain whether a repeat biopsy should be performed in men with clinically localized PCa and prior negative biopsy findings. In men with a negative first biopsy but persistently high PSA, the European Association of Urology (EAU) [75] guidelines recommend a prostate biopsy; however, among men with suspicion of having PCa and a prior negative biopsy, a repeat biopsy was reported negative in approximately 80% of the men. In addition to being costly, biopsies are associated with morbidity, anxiety, discomfort and complications [132]. New biomarkers may increase the diagnostic accuracy of repeat biopsies, and reduce the number of unnecessary biopsies, but the long term health outcomes are unclear.

Results of recent studies have shown the potential clinical utility of the urine based PROGENSA prostate cancer antigen 3 assay (PCA3) to predict repeat biopsy outcomes in men with elevated serum PSA levels and previous negative biopsy findings [7, 8, 43, 47, 50, 65, 69, 105, 126, 136], and reported that an increasing PCA3 score corresponds to an increasing probability of a positive repeat biopsy. The PCA3 test has been shown in some studies to be superior to serum PSA in predicting biopsy outcome [60, 64, 105] and has been included in recently developed nomograms [10, 42, 122]. A recent literature review reported that current evidence suggests PCA3 is clinically useful for selecting which patients should have a repeat biopsy [59]. Several studies have found that TMPRSS2:ERG assay (T2:ERG) is also associated with biopsy outcome [44, 77, 92, 94, 140, 151, 167], and may better discriminate between low and high grade cancers [167]. Although there are studies supporting increased diagnostic accuracy for both biomarkers, the ideal thresholds to trigger a repeat biopsy, and the resulting increase in survival and decrease in unnecessary

biopsies are unknown.

In this chapter, we use decision analysis to evaluate the clinical value of PCA3 and T2:ERG in men with clinically localized PCa who had at least one prior negative biopsy. We perform head-to-head comparisons of protocols that use either PCA3 or T2:ERG in terms of the incremental change in 10-year overall survival and the rate of negative biopsies. Furthermore, we consider 15-year cancer-specific survival as an end point in our analyses. We present results for both expected 10-year overall survival and 15-year cancer-specific survival and the rate of repeat biopsy for each biomarker. We also present results of sensitivity analysis using Monte Carlo sampling of clinical factors such as PSA level, biopsy detection rate and age to provide evidence about which patients benefit most from the use of an additional biomarker.

The remainder of this chapter is organized as follows. In Section 4.2, we describe the decision model of the problem and the approach for probabilistic sensitivity analyses, and provide a review of the relevant literature on the survival estimates. In Section 4.3, we present results from statistical analyses and probabilistic sensitivity analyses addressing the impact of parameter uncertainty on the model outcomes. Finally, in Section 4.4, we highlight main conclusions and limitations of this study.

4.2. Model and Methods

We employ a decision analytic framework using Monte Carlo simulation and statistical modeling to develop testing protocols integrating PSA with PCA3 and T2:ERG. We conduct sensitivity analysis around model parameters to assess the impact of parameter uncertainty on model outcomes.

4.2.1. Study Population

The decision analysis model was based on the results from a prospectively collected cohort design. For the study cohort, post-DRE urine was prospectively collected from $1,977$ men presenting for diagnostic prostate biopsy at three U.S. academic institutions ($n = 733$) and 7 community clinics ($n = 1,244$). The vast majority of men had elevated serum PSA. As this cohort reflects actual clinical practice, no specific indication for repeat biopsy was required; however the vast majority was for persistently elevated serum PSA. Exclusion criteria included the following: prior attempted curative therapy (radical prostatectomy

(RP), radiation therapy (RT), androgen deprivation therapy (ADT) or brachytherapy), surgical treatment of the prostate within 6 months of urine collection (or previous biopsy within 6 weeks), taking 5a-reductase inhibitors or testosterone within 3 months of urine collection, or prostatitis at the time of urine collection. All urine specimens were obtained with institutional review board approval.

4.2.2. Decision Tree

We constructed a decision tree to compare the expected 10-year survival and and 15-year cancer-specific survival for protocols that use one of the urinary biomarkers versus those that do not for patients with elevated PSA. The complete decision tree schema is shown in Figure 4.1. The initial decision is whether to use an additional biomarker (yes or no) or no repeat biopsy; therefore the decision tree branches out to three separate decision arms. *Branch 1* represents the protocols that incorporate a urinary biomarker into repeat biopsy decisions. *Branch 2* represents the protocol that does not involve any additional indication for repeat biopsy, therefore every patient is assumed to receive biopsy regardless of his clinical parameters (age, serum PSA level etc.). *Branch 3* represents the protocol that no patient receives a repeat biopsy.

In the decision tree, men with detected and undetected clinically localized PCa were assumed to have 10-year survival consistent with men who receive RP at diagnosis and men under conservative treatment (whose cases were managed without surgery or radiation,), respectively. The risk of PCa was derived from the PCPTRC version 2.0 risk calculator, which incorporates age, race, PSA level, family history of PCa, DRE and history of a negative biopsy [120]. The decision tree accounts for the different cancer grades based on patients Gleason score (GS) (GS$<$ 7, GS$=$ 7, and GS$>$ 7). The probability for each grade was estimated based on the proportion of each outcome in the study population.

The biopsy decision in *Branch 1* of the decision tree is determined by a pre-specified threshold for the urinary biomarker. The probability that the biomarker score exceeds this threshold is grade dependent and estimated from the study population (See Table 4.2). The probability of a positive repeat biopsy was estimated from Haas et al. [68]. The primary end point of each branch is the 10-year overall survival estimated from [159], which depends on cancer grade, age, serum PSA level, race and Charlson comorbidity index (CCI). We did not have CCI for the patients included in our study; thus, we assumed that they are healthy patients with CCI in the range of $0-1$. Outcomes for patients without PCa were

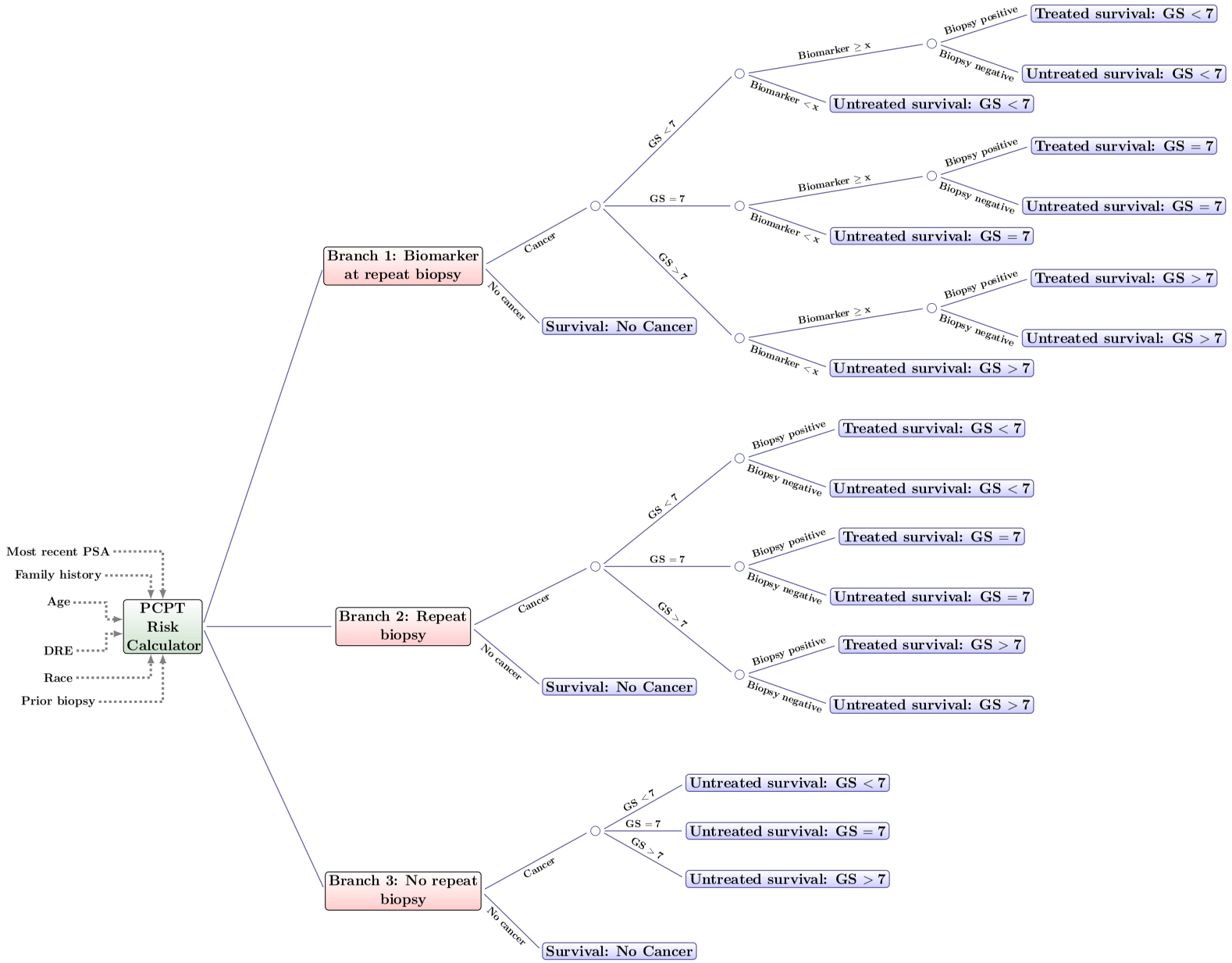

Figure 4.1: Decision tree schema for repeat biopsy decisions based on biomarker tests.

also taken from Tewari et al. [159]. The look-up tables for 10-year overall survival are constructed separately for black and white males; however, Tewari et al. [159] did not find race to be an independent predictor of survival, and most of the patients in our study population were white. Thus, we considered the 10-year overall survival estimates in white males with clinically localized PCa.

We conducted similar analyses using the 15-year prostate cancer-specific survival as the primary end point in the decision tree. We used 15-year cancer-specific survival since there does not exist a long follow-up study in literature that estimates 15-year overall survival. We obtained the 15-year cancer-specific survival estimates for untreated, clinically localized PCa from Johansson et al. [85], and the 15-year cancer-specific survival estimates after RP from Stephenson et al. [152].

There is no consensus about the most appropriate threshold for the PCA3 and T2:ERG tests. The FDA recommends a PCA3 threshold of 25, but a threshold of 35 is commonly used [47, 50, 69, 76, 105, 135, 146, 174]. While some studies have found the cutoff of 25 provides a good balance between sensitivity and specificity [9, 116, 125, 157], others have supported the use of different thresholds, e.g. 17 [10, 42], 43 [64] and 51 [122]. In this study we consider the thresholds of 25, 35 and 100 for PCA3. Regarding the T2:ERG threshold, Tomlins et al. [167] considered the specimens with T2:ERG score > 50 as positive, and Leyten et al. [94] considered the threshold of 10 in their multivariate regression analysis. In this study, to provide a diverse set of thresholds, we considered the thresholds of $7, 10, 30, 50$ and 100.

4.2.3. Survival Estimates

We conducted a literature review to obtain estimates of overall survival in clinically localized PCa patients. Some relevant studies were based on retrospective cohorts that did not receive PSA screening. One important question about these studies is how the findings from pre-PSA era relate to the current era when many cancers are detected by PSA testing. Sweat et al. [156] estimated long-term (20-year) probability of death from prostate cancer and other competing causes stratified by age and GS based on the study population diagnosed with clinically localized PCa between 1971 and 1984. The estimates provided in Albertsen et al. [4] were also derived from a retrospective cohort of patients diagnosed between 1971 and 1976. The estimates from these studies do not incorporate the effect of the lead-time associated with the screening for PSA, and as a consequence, they underestimate

long-term survival outcomes in contemporary settings.

The study by Liu et al. [99] evaluated the effectiveness of the RP compared to other treatments in older men with local/regional PCa and CCI ≤ 1 based on a relatively large population identified from Surveillance, Epidemiology, and End Results (SEER) and Medicare linked data. The study provided all-cause and cancer specific 5- and 10-year survival rates for patients receiving RP versus other treatment modalities but not stratified by clinical parameters of men with local/regional PCa. More recently, Kibel et al. [87] investigated the differences in overall-survival in men with clinically localized PCa treated with RP or radiation therapy. Both studies by Liu et al. [99] and Kibel et al. [87] did not provide overall grade-specific survival for patients with clinically localized PCa.

There are several studies that developed nomograms to predict the overall survival in clinically localized PCa. For the patients included in our study, we did not have the complete information for the 11 clinical variables required for the Cowen et al. [45] nomogram. The nomogram developed by Walz et al. [173] has the advantage of being based entirely on patients from the PSA screening era, and only requires two clinical parameters (age and CCI) to predict the probability of 10-year life expectancy after RP or RT; however, the probabilities do not account for the available clinicopathological information. Albertsen et al. [4] has the benefit of having outcomes for approximately $20,000$ men who were treated with conservative management after the diagnosis of the localized PCa, but the study considers men older than 66 years, and 38% of our study population is under age 66 years. Due to the shortcomings of the above referenced studies, we used the overall 10-year survival estimates from Tewari et al. [159], which quantify the impact of treatment modality on overall-survival of men with clinically localized PCa.

We further performed sensitivity analyses using the 15-year survival estimates obtained from literature. We could not find 15-year grade-specific overall survival for cancerous and cancer-free patients. Therefore, we used the 15-year cancer-specific mortality rates after RP from Stephenson et al. [152] and 15-year cancer-specific survival for clinically localized untreated PCa from Johansson et al. [85]. Additional assumptions related to the study design of these two studies for 15-year cancer-specific survival are as follows:

15*-year survival for clinically localized, untreated PCa from Johansson et al.* [85]:

1. The study cohort in Johansson et al. [85] included patients with early, initially untreated PCa. PSA testing was not available and no screening activities for PCa

took place during the period when the cohort was recruited. Thus, the survival estimates from this study systematically underestimate the 15-year survival for untreated clinically localized PCa diagnosed during the PSA era.

2. In this study, the World Health Organization (WHO) classification of malignant diseases was used. It is mentioned in the study that the grades based WHO classification system are not directly translatable to the Gleason grading system. However, based on the report [41], grade 1 was compared with GS 2 to 4, grade 2 with GS 5 to 7, and grade 3 with GS 8 to 10. Using this conversion, we had to make the assumption that the 15-year untreated cancer-specific survival rates for GS < 7 and GS $= 7$ cancers are the same.

3. Patients were given no initial treatment if the tumor growth was localized to the prostate gland as judged by DRE and no distant metastases were present. Patients 75 years or older were not included in the study. Clinical examination, laboratory tests, and bone scans were performed every 6 months during the first 2 years after diagnosis and subsequently once a year during the first 10 years of observation and thereafter at least once every second year. Patients in whom the cancer progressed to symptomatic disease were treated with exogenous estrogens or orchidectomy.

15-*year cancer-specific mortality after RP from Stephenson et al.*[152]:

1. The study cohort in Stephenson et al. [152] included patients who underwent RP for localized PCa between 1987 and 2005 to construct a nomogram for patients treated in the era of PSA screening.

2. Patients were observed for disease recurrence after biopsy with regular serum PSA determinations and clinical assessment at 3- to 6-month intervals for the first 5 years and annually thereafter.

4.2.4. Probabilistic Sensitivity Analyses

We conducted multi-way probabilistic sensitivity analyses around model parameters (biopsy sensitivity, sensitivity of biomarkers at different thresholds for different cancer grades and 10-year survival under different treatments) and clinical parameters (serum PSA and age) using Monte Carlo simulation. We did not conduct multi-way sensitivity analyses representing the uncertainty around clinical parameters in the analysis of 15-year cancer-specific

survival since the 15-year survival estimates are not given by PSA and age. Based on the sample distribution, we assumed a gamma distribution for PSA and age. The uncertainty around the other probabilities was represented using beta distributions. Parameters alpha and beta were derived from the mean (μ) and standard deviation (σ) using the following formulas: $\alpha = \mu^2 \times (1-\mu)/\sigma^2$ and $\beta = \mu \times (1-\mu)/\sigma^2 - \alpha$. Table C.1 shows the assumed distributions and the corresponding distribution parameters for the model probabilities and patient parameters that were used in the probabilistic sensitivity analysis. Tables C.2 and C.3 show the assumed distributions and the corresponding distribution parameters parameters for overall 10-year survivals and 15-year cancer-specific survivals used in probabilistic sensitivity analysis. We sampled the parameters $1,000$ times drawn from independent distributions and compute the additional 10-year survival and percentage of men biopsied for each resulting decision tree. We chose 4 and 30 ng/ml, and 50 and 75 years as lower and upper bounds on serum PSA and age, respectively.

4.3. Results

In this section, we present numerical results from our analyses on base case and address the effect of parameter uncertainty through probabilistic sensitivity analyses.

4.3.1. Specimen Collection and Processing: T2:ERG and PCA3 Assay

Urine processing for determination of PCA3 and T2:ERG scores was performed as described in prior studies [44, 167, 179]. Urine specimens were obtained immediately after attentive DRE, refrigerated, and processed within 4 hours by mixing with an equal volume of urine transport medium and stored below 70°C until analysis. Amounts of urine PCA3, T2:ERG and PSA mRNA were determined with transcription mediated amplification (TMA) assays. To generate a T2:ERG score, the amount of T2:ERG mRNA is normalized to the amount of PSA mRNA, which is calculated using the following formula: ($100,000\times$ average urine T2:ERG copies/mL) / (average urine PSA copies/mL). Samples with average urine PSA copies/mL of more than $10,000$ copies/mL were considered informative for urine T2:ERG scores. Urine T2:ERG scores were assessed as described using the final T2:ERG TMA assay as described in [44, 151, 179] or an earlier generation assay 22 that yield equivalent T2:ERG scores.

The PROGENSA PCA3 assay similarly quantitates PCA3 and PSA mRNA in post-

DRE urine. The PCA3 score was calculated with the following formula: $1,000\times$ (average urine PCA3 copies/mL) / (average urine PSA copies/mL). Samples with average urine PSA copies/mL of more than $10,000$ copies/mL were considered informative. Identical primers for quantifying urine PSA are used in the PROGENSA PCA3 assay and T2:ERG assay.

All urine PCA3 and T2:ERG analysis was performed at the University of Michigan or Gen-Probe, Inc, with a subset of samples assessed at both to ensure concordance. A total of $1,936$ urine samples had sufficient urine PSA ($> 10,000$ copies/mL) to provide informative PCA3 and T2:ERG scores, and these samples were considered for analysis. The final study population consisted of 140 men with informative urine PCA3 and T2:ERG scores who had a history of at least one negative previous biopsy and were diagnosed with prostate cancer in their study biopsy.

4.3.2. Study Population

Table 4.1 provides the characteristics of the 420 men who had previous negative biopsies. Men with a positive biopsy had a statistically significant higher age, lower prostate volume, and a higher mean PCA3 and T2:ERG scores than men with a negative repeat biopsy. Mean serum PSA did not significantly change between men with negative versus positive biopsy. Men with a positive biopsy had clinical stage T1 and T2 in 78.6% and 20% of cases, respectively; 88.6% had a biopsy GS of $6-7$ and 75% had $\leq 33\%$ positive cores.

Table 4.2 provides our estimates of the probability that a man's biomarker scores exceed different thresholds, based on the man's grade of PCa. Among 420 men with stage T1 or T2 prostate cancer, 140 (33.3%) had cancer on repeat biopsy. Of the 140 men with positive repeat biopsy, 82 (58.6%) had GS < 7 cancer, 42 (30.0%) had GS $= 7$ and 16 (11.4%) had GS > 7 cancer. Based on univariate analysis, all of the pre-biopsy clinical variables were associated with a positive repeat biopsy ($p < 0.04$) (data not shown).

The performance characteristics at different PSA, PCA3 and T2:ERG thresholds are presented in Table 4.3. The PCA3 threshold of 25 demonstrated the highest accuracy in predicting the positive repeat biopsy (AUC : 0.652), compared to the PSA threshold of 10 (AUC : 0.54) in Table 4.3. The PCA3 threshold of 35 provided a good balance between sensitivity (49.3%) and specificity (74.3%). In comparison, the sensitivity and specificity of T2:ERG with a biopsy threshold of 10 were 50.7% and 68.6%, respectively. Considering the test outcomes as continuous variables, the AUC was 0.687 for the PCA3 score ($p < 0.0001$),

Table 4.1: Baseline characteristics of the study population.

	Men with negative biopsy (n = 280)	Men with positive biopsy (n = 140)	p - value	All men (n = 420)
	Mean ± SD / number (%)	Mean ± SD / number (%)		Mean ± SD / number (%)
Age (years)	65.4 ± 8.1	68.2 ± 8.9	0.002*	66.3 ± 8.5
Serum PSA (ng/ml)	7.2 ± 5.4	8.4 ± 6.5	0.0766†	7.5 ± 5.8
Men with serum PSA (ng/ml) (%)				
< 4	55 (19.6)	24 (17.1)		79 (18.8)
$4 - 10$	177 (63.2)	83 (59.3)		260 (61.9)
> 10	48 (17.1)	33 (23.6)		81 (19.3)
No. Ethnicity (%)			0.0113‡	
African-American	12 (4.3)	15 (10.7)		27 (6.4)
Other	268 (95.7)	125 (89.3)		393 (93.6)
No. DRE result			0.0783‡	
Normal	242 (86.4)	110 (78.6)		352 (83.8)
Abnormal	33 (11.8)	28 (20.0)		61 (14.5)
Not available	5 (1.8)	2 (1.4)		7 (1.7)
Prostate volume ($n = 273/136/409$)	69.3 ± 39.8	58.2 ± 33.1	0.0016†	65.6 ± 38.0
PCA3 score	32 ± 36.4	61 ± 78.4	< 0.0001†	41.7 ± 55.8
T2:ERG score	32.8 ± 110.4	127.5 ± 678.3	0.0006†	64.4 ± 403.4

Abbreviations: SD, Standard deviation; PSA, Prostate-specific antigen; DRE, digital rectal examination; PCA3, Prostate Cancer Antigen 3, T2:ERG, the transmembrane protease, serine 2 (TMPRSS2): v-ets erythroblastosis virus E26 oncogene homolog (avian) (ERG) fusion. *t-test, † Wilcoxon rank sum test, ‡χ^2 test.

0.602 for T2:ERG ($p = 0.0273$) and 0.553 for serum PSA ($p = 0.052$).

Table 4.4 shows PCa detection rates for varying PSA, PCA3 and T2:ERG thresholds, the number of prostate biopsies that would be avoided and PCa cases with GS ≥ 7 that would be missed if the urinary biomarker (PCA3 or T2:ERG) was used to select men for repeat biopsies. A PCA3 threshold ≥ 25 and ≥ 35 would detect 95 (67.9%) and 69 (49.3%) of PCa cases, respectively. A T2:ERG threshold ≥ 7 and ≥ 10 showed similar performance detecting 78 (55.7%) and 71 (50.7%) of PCa cases, respectively. A PCA3 threshold ≥ 25 would identify 42 (72.4%) of 58 cancer cases with GS ≥ 7 and avoid 52.4% of repeat

Table 4.2: Probability estimates for the biomarkers PCA3 and T2:ERG at different thresholds.

Probability	No. of Patients, (%)	Probability	No. of Patients, (%)	Probability	No. of Patients, (%)
$\mathbb{P}(\text{PCA3} \geq 25 \mid \text{GS} < 7)$	53/82 (64.6)	$\mathbb{P}(\text{T2ERG} \geq 7 \mid \text{GS} < 7)$	43/82 (52.4)	$\mathbb{P}(\text{T2ERG} \geq 50 \mid \text{GS} < 7)$	21/82 (25.6)
$\mathbb{P}(\text{PCA3} \geq 25 \mid \text{GS} = 7)$	31/42 (73.8)	$\mathbb{P}(\text{T2ERG} \geq 7 \mid \text{GS} = 7)$	23/42 (54.8)	$\mathbb{P}(\text{T2ERG} \geq 50 \mid \text{GS} = 7)$	11/42 (26.2)
$\mathbb{P}(\text{PCA3} \geq 25 \mid \text{GS} > 7)$	11/16 (68.8)	$\mathbb{P}(\text{T2ERG} \geq 7 \mid \text{GS} > 7)$	12/16 (75.0)	$\mathbb{P}(\text{T2ERG} \geq 50 \mid \text{GS} > 7)$	8/16 (50.0)
$\mathbb{P}(\text{PCA3} \geq 35 \mid \text{GS} < 7)$	37/82 (45.1)	$\mathbb{P}(\text{T2ERG} \geq 10 \mid \text{GS} < 7)$	38/82 (46.3)	$\mathbb{P}(\text{T2ERG} \geq 100 \mid \text{GS} < 7)$	13/82 (15.9)
$\mathbb{P}(\text{PCA3} \geq 35 \mid \text{GS} = 7)$	23/42 (54.8)	$\mathbb{P}(\text{T2ERG} \geq 10 \mid \text{GS} = 7)$	21/42 (50.0)	$\mathbb{P}(\text{T2ERG} \geq 100 \mid \text{GS} = 7)$	7/42 (16.7)
$\mathbb{P}(\text{PCA3} \geq 35 \mid \text{GS} > 7)$	9/16 (56.3)	$\mathbb{P}(\text{T2ERG} \geq 10 \mid \text{GS} > 7)$	12/16 (75.0)	$\mathbb{P}(\text{T2ERG} \geq 100 \mid \text{GS} > 7)$	6/16 (37.5)
$\mathbb{P}(\text{PCA3} \geq 100 \mid \text{GS} < 7)$	9/82 (11.0)	$\mathbb{P}(\text{T2ERG} \geq 30 \mid \text{GS} < 7)$	30/82 (36.6)		
$\mathbb{P}(\text{PCA3} \geq 100 \mid \text{GS} = 7)$	10/42 (23.8)	$\mathbb{P}(\text{T2ERG} \geq 30 \mid \text{GS} = 7)$	12/42 (28.6)		
$\mathbb{P}(\text{PCA3} \geq 100 \mid \text{GS} > 7)$	4/16 (25.0)	$\mathbb{P}(\text{T2ERG} \geq 30 \mid \text{GS} > 7)$	8/16 (50.0)		

Table 4.3: Performance of serum PSA, PCA3, and T2:ERG in predicting PCa at repeat biopsies: Univariate analyses.

Biomarkers	Sensitivity, (%)	Specificity, (%)	PPV, (%)	NPV, (%)	AUC (95% CI)	p-value
PSA ≥ 10	24.3	56.8	44.5	78.7	$0.54\ (0.49 - 0.58)$	0.083
PCA3 score						
≥ 25	67.9	62.5	47.5	79.6	$0.65\ (0.60 - 0.70)$	< 0.0001
≥ 35	49.3	74.3	48.9	74.6	$0.62\ (0.57 - 0.67)$	< 0.0001
T2:ERG						
≥ 7	55.7	62.1	42.4	73.7	$0.59\ (0.54 - 0.64)$	0.0006
≥ 10	50.7	68.6	44.7	73.6	$0.60\ (0.55 - 0.65)$	0.0001
≥ 30	35.7	80.0	47.2	71.3	$0.58\ (0.53 - 0.63)$	0.0006
≥ 100	18.6	92.5	55.3	69.4	$0.56\ (0.52 - 0.59)$	0.001

Abbreviations: PPV, Positive predictive value; NPV, Negative predictive value.

biopsies, and a threshold of ≥ 35 would identify 32 (55.2%) cancer cases with GS ≥ 7, but 66.4% of all biopsies could have been avoided. Similarly, a T2:ERG threshold ≥ 7 would identify 35 (60.3%) of 58 cancer cases with GS ≥ 7 and avoid 56.2% of repeat biopsies, and a threshold of ≥ 10 would identify 33 (56.9%) cancer cases with GS ≥ 7, but 62.1% of all biopsies could have been avoided.

4.3.3. Base Case Analysis

We considered a base case patient with the following characteristics: white, age 65 years, most recent serum PSA of 6.3 ng/ml based on the mean PSA of patients in the study cohort, CCI of 0, no family history of prostate cancer, normal DRE and a prior negative biopsy. These attributes were chosen to represent an average patient in the study population. Table 4.5 presents 10-year survival and biopsy rates for the protocols with varying biopsy thresholds. Table 4.5 shows that *Branch 2* (repeat biopsy) yields better 10-year survival than *Branch 1* (biomarker at repeat biopsy) under every protocol with various PCA3 and T2:ERG thresholds. Similar results were obtained in the analysis of 15-year cancer-specific survival (see Table 4.6).

4.3.4. Probabilistic Sensitivity Analyses

Multi-way probabilistic sensitivity analyses consisted of two steps. The first step involved varying the model parameters. The results summarized in Table 4.7 show that the confi-

Table 4.4: PCa detection with varying PSA, PCA3 and T2:ERG thresholds for repeat biopsy.

	Biopsied men, No. (%)	**PCa cases, No. (%)**	**PPV, (%)**	**Missed PCa , No. (%) (n =140)**	**Prostate cancers GS $\geq$ 7 missed, No. (%) (n =58)**	**Biopsies avoided, No. (%) (n =420)**
No threshold	420	140	33.3	–	–	–
PSA, ng/ml						
$\geq$ 2.5	387 (92.1)	134 (95.7)	34.6	6 (4.3)	1 (1.7)	33 (7.9)
$\geq$ 4.0	341 (81.2)	116 (82.9)	34.0	24 (17.1)	7 (12.1)	79 (18.8)
$\geq$ 10.0	82 (19.5)	34 (24.3)	41.5	106 (75.7)	40 (69.0)	338 (80.5)
PCA3 score						
$\geq$ 25	200 (47.6)	95 (67.9)	47.5	45 (32.1)	16 (27.6)	220 (52.4)
$\geq$ 35	141 (33.6)	69 (49.3)	48.9	71 (50.7)	26 (44.8)	279 (66.4)
$\geq$ 100	42 (10.0)	23 (16.4)	54.8	117 (83.6)	44 (75.9)	378 (90.0)
T2:ERG						
$\geq$ 7.0	184 (43.8)	78 (55.7)	42.4	62 (44.3)	23 (39.7)	236 (56.2)
$\geq$ 10.0	159 (37.9)	71 (50.7)	44.7	69 (49.3)	25 (43.1)	261 (62.1)
$\geq$ 30.0	106 (25.2)	50 (35.7)	47.2	90 (64.3)	38 (65.5)	314 (74.8)
$\geq$ 50.0	86 (61.4)	40 (28.6)	46.5	100 (71.4)	39 (67.2)	334 (79.5)
$\geq$ 100.0	47 (11.2)	26 (18.6)	55.3	114 (81.4)	45 (77.6)	373 (88.8)

Table 4.5: 10-year life survival and percentage of men biopsied for the base case patient at various biopsy thresholds for PCA3 and T2:ERG.

Biomarkers at different thresholds	**Percentage of men biopsied at this threshold**	**10-year survival**	**Percentage change in survival***
Branch 1†			
PCA3			
$\geq$ 25	44.6	83.98	0.93 (0.66, 1.14)
$\geq$ 35	31.2	83.44	1.47 (1.04, 1.78)
T2:ERG			
$\geq$ 7	41.5	83.63	1.27 (0.91, 1.56)
$\geq$ 10	35.3	83.50	1.41 (1.00, 1.73)
$\geq$ 30	23.1	83.03	1.88 (1.33, 2.30)
$\geq$ 50	18.6	83.84	2.07 (1.47, 2.53)
$\geq$ 100	9.6	82.54	2.36 (1.68, 2.90)
Branch 2‡	23.5	84.91	–

*This is the absolute difference between *Branch 1* and *Branch 2*. The numbers in the parentheses are calculated using the 95% Confidence Intervals estimated by Tewari et al. [159] on the 10-year survival under radical prostatectomy and conservative management; †*Branch 1* uses urinary biomarkers at repeat biopsy; ‡*Branch 2* has no indication at repeat biopsy.

Table 4.6: 15-year cancer-specific life survival and percentage of men biopsied for the base case patient at various biopsy thresholds for PCA3 and T2:ERG.

Biomarkers at different thresholds	Percentage of men biopsied at this threshold	15-year survival	Percentage change in survival[§]	Percentage change in survival[¶]
Branch 1*				
PCA3				
≥ 25	44.63	85.63	1.67 (0.80, 2.47)	3.42 (1.65, 5.02)
≥ 35	31.25	84.70	2.60 (1.25, 3.84)	2.49 (1.19, 3.64)
T2:ERG				
≥ 7	42.05	85.08	2.22 (1.08, 3.29)	2.87 (1.36, 4.19)
≥ 10	35.95	84.83	2.47 (1.20, 3.66)	2.62 (1.24, 3.82)
≥ 30	23.70	84.10	3.20 (1.56, 4.72)	1.89 (0.89, 2.76)
≥ 50	19.28	83.71	3.59 (1.74, 5.30)	1.51 (0.70, 2.18)
≥ 100	10.11	83.20	4.10 (1.99, 6.06)	0.99 (0.46, 1.43)
Branch 2[†]	100.0	87.30	–	–
Branch 3[‡]	0.0	82.21	–	–

The numbers in the parentheses are calculated using the 95% confidence interval (CI)s for 15-year cancer-specific mortality after and survival for untreated clinically localized PCa given in studies by Stephenson et al. [152] and Johansson et al. [85], respectively; **Branch 1* uses urinary biomarkers at repeat biopsy; [†]*Branch 2* has no indication at repeat biopsy; [‡]*Branch 3* is no repeat biopsy; [§]The difference is calculated as the absolute difference between *Branch 1* and *Branch 2* in the decision tree; [¶]The difference is calculated as the absolute difference between *Branch 1* and *Branch 3* in the decision tree.

dence interval for each protocol is relatively narrow, and the magnitude of effect difference for each protocol was not changed when uncertainty was incorporated for the base case patient. In the second step, we performed a sensitivity analysis including the uncertainty around serum PSA level and age of the base case patient in addition to varying the model parameters (Table 4.8). Multi-way sensitivity analyses demonstrated that *Branch 2* (repeat biopsy) yields better 10-year survival than *Branch 1* (biomarker at repeat biopsy) under every protocol with various PCA3 and T2:ERG thresholds. Similar results were obtained in the analysis of 15-year cancer-specific survival (See Table 4.9).

Figure 4.2 shows the results from multi-way probabilistic sensitivity analyses on the base case. The y-axis shows the absolute change in 10-year survival between the decision arms represented by *Branch 1* and *2* in the decision tree, and the x-axis shows the percentage of men biopsied under the protocols that use PCA3 and T2:ERG at different thresholds. In this analysis, the age and serum PSA of the base case remained constant; however, we represented the uncertainty around the other model parameters with the distributions described in Tables 4.6 and 4.9. We can see from this figure that the use of PCA3 test with a threshold of 25 at repeat biopsy provides expected 10-year overall survival close to the case where there is no indication for repeat biopsy. Using T2:ERG assay would cause

Table 4.7: Multi-way probabilistic sensitivity analysis representing the uncertainty around model parameters.

Biomarkers at different thresholds	Percentage of men biopsied at this threshold , (95% CI)	10-year survival	Percentage change in survival*, (95% CI)
Branch 1†			
PCA3			
≥ 25	45.65 (45.12 – 46.19)	83.91 (83.83 – 84.0)	0.92 (0.91 – 0.94)
≥ 35	32.33 (31.89 – 32.33)	83.38 (83.30 – 83.47)	1.45 (1.43 – 1.48)
T2:ERG			
≥ 7	42.85 (42.31 – 43.39)	83.57 (83.48 – 83.65)	1.27 (1.25 – 1.29)
≥ 10	36.82 (36.35 – 37.30)	83.43 (83.35 – 83.52)	1.40 (1.38 – 1.42)
≥ 30	24.70 (24.34 – 25.05)	82.97 (82.89 – 83.05)	1.87 (1.84 – 1.89)
≥ 50	22.22 (21.89 – 22.56)	82.79 (82.71 – 82.87)	2.05 (2.02 – 2.07)
≥ 100	8.62 (8.42 – 8.82)	82.50 (82.41 – 82.58)	2.34 (2.31 – 2.37)
Branch 2‡	100.0	84.84 (84.75 – 84.93)	–

*The difference is calculated as the absolute difference between *Branch 1* and *Branch 2* in the decision tree. †*Branch 1* uses urinary biomarkers at repeat biopsy; ‡*Branch 2* has no indication at repeat biopsy.

Table 4.8: Multi-way probabilistic sensitivity analyses representing the uncertainty around model and clinical parameters (serum PSA and age).

Biomarkers at different thresholds	Percentage of men biopsied at this threshold, (95% CI)	10-year survival	Percentage change in survival*, (95% CI)
Branch 1†			
PCA3			
≥ 25	46.80 (46.27 – 47.32)	82.16 (81.72 – 82.61)	1.36 (1.31 – 1.42)
≥ 35	33.53 (33.10 – 33.96)	81.38 (80.92 – 81.85)	2.14 (2.22 – 2.06)
T2:ERG			
≥ 7	43.55 (43.04 – 44.07)	81.65 (81.19 – 82.11)	1.88 (1.81 – 1.95)
≥ 10	37.99 (37.51 – 38.47)	81.44 (80.98 – 81.91)	2.09 (2.01 – 2.17)
≥ 30	25.47 (25.12 – 25.83)	80.77 (80.29 – 81.26)	2.75 (2.65 – 2.86)
≥ 50	22.66 (22.34 – 22.98)	80.49 (80.0 – 80.98)	3.04 (2.92 – 3.15)
≥ 100	9.34 (9.13 – 9.54)	80.07 (79.57 – 80.57)	3.46 (3.59 – 3.33)
Branch 2‡	100.0	83.53 (83.12 – 83.94)	–

*The difference is calculated as the absolute difference between *Branch 1* and *Branch 2* in the decision tree. †*Branch 1* uses urinary biomarkers at repeat biopsy; ‡*Branch 2* has no indication at repeat biopsy;

Table 4.9: Multi-way probabilistic sensitivity analysis for 15-year cancer-specific survival representing the uncertainty around model parameters.

Biomarkers at different thresholds	Percentage of men biopsied at this threshold, (95% CI)	15-year survival	Percentage change in survival[§], (95% CI)	Percentage change in survival[¶], (95% CI)
Branch 1*				
PCA3				
≥ 25	45.46 (44.91 – 46.01)	85.57 (85.37 – 85.77)	1.47 (1.44 – 1.51)	2.34 (2.26 – 2.42)
≥ 35	32.36 (31.92 – 32.79)	84.64 (84.44 – 84.84)	2.16 (2.10 – 2.23)	2.11 (2.05 – 2.17)
T2:ERG				
≥ 7	42.82 (42.29 – 43.35)	85.03 (84.83 – 85.23)	2.21 (2.17 – 2.25)	2.87 (2.82 – 2.92)
≥ 10	36.79 (36.35 – 37.30)	84.77 (84.57 – 84.97)	2.47 (2.42 – 2.52)	2.61 (2.56 – 2.65)
≥ 30	24.59 (24.23 – 24.95)	84.05 (83.85 – 84.25)	3.17 (3.13 – 3.24)	1.89 (1.86 – 1.93)
≥ 50	21.93 (21.60 – 22.27)	83.68 (83.48 – 83.89)	3.55 (3.49 – 3.62)	1.53 (1.50 – 1.55)
≥ 100	8.54 (8.74 – 8.35)	83.16 (82.95 – 83.36)	4.08 (4.01 – 4.15)	1.00 (0.98 – 1.02)
Branch 2†	100.0	87.24 (87.04 – 87.44)	–	–
Branch 3‡	0.0	82.16 (81.95 – 82.37)	–	–

*_Branch 1_ uses urinary biomarkers at repeat biopsy; †_Branch 2_ has no indication at repeat biopsy; ‡_Branch 3_ is no repeat biopsy; §The difference is calculated as the absolute difference between _Branch 1_ and _Branch 2_ in the decision tree; ¶The difference is calculated as the absolute difference between _Branch 1_ and _Branch 3_ in the decision tree.

fewer biopsies; however, the additional 10-year survival would be less than the case where PCA3 with a threshold of 25.

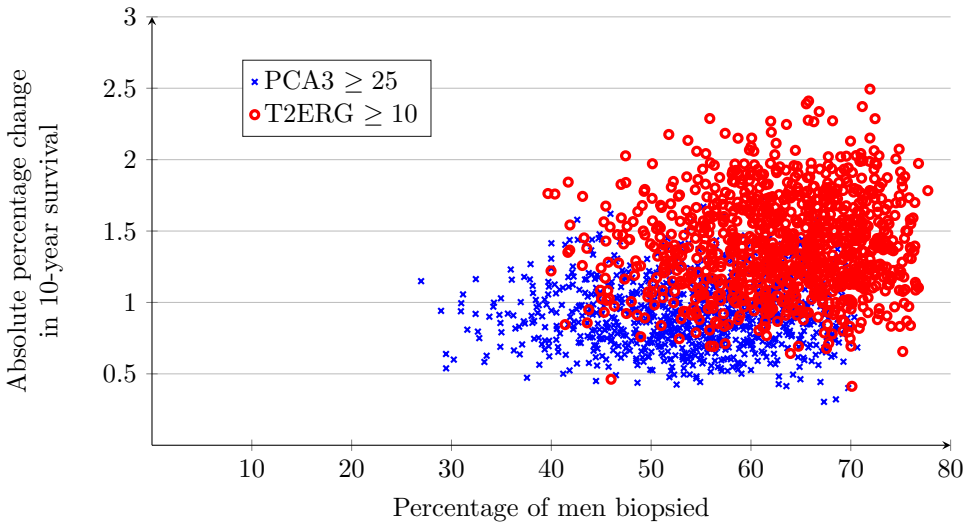

Figure 4.2: Multi-way sensitivity analyses on the base case patient with fixed age and PSA.

Figure 4.3 demonstrates the results from multi-way probabilistic sensitivity analyses. In this analysis, we represented the uncertainty around age, serum PSA and model parameters with probability distributions described in Tables 4.6 and 4.9. We can see from this figure that the use of PCA3 test with a threshold of 25 and T2:ERG of 10 at repeat biopsy

decisions provides expected 10-year overall survival close to the case where there is no indication for repeat biopsy. In summary, T2:ERG and PCA3 perform similarly once we incorporated uncertainty around serum PSA and age.

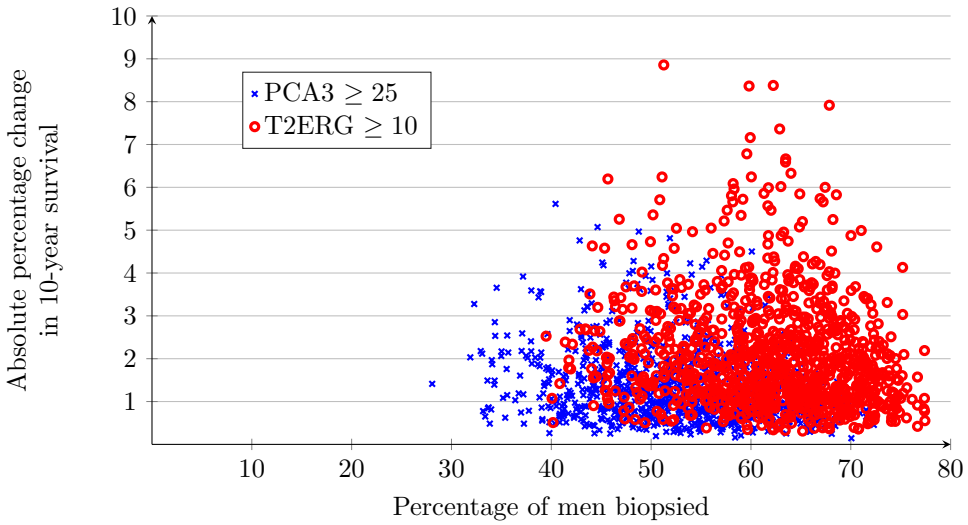

Figure 4.3: Multi-way sensitivity analyses with variation in age and PSA.

4.4. Conclusions

There is no definitive criterion to decide whether to perform a repeat prostate biopsy. Typically the decision to perform a repeat biopsy is based on the measurement of serum PSA and the findings of a DRE. The use of diagnostic biomarkers such as PCA3 and T2:ERG may help clinicians make better decisions about repeat biopsies. In this respect, the PCA3 assay and T2:ERG have shown promising results, and the studies available in the literature support the use of PCA3 in patients with persistent suspicions of PCa who had undergone previous negative biopsies. However, these studies focus on diagnostic performance and not health outcomes. In this study, we investigated the value of PCA3 and T2:ERG for improving the overall 10-survival and reducing unnecessary repeat biopsies in the challenging subgroup of patients with previous negative biopsies and persistently elevated PSA levels.

Based on multi-way sensitivity analysis for the base case patient, the protocols using a PCA3 threshold of ≥ 25 and a T2:ERG threshold of ≥ 10 at repeat biopsy decisions resulted in a 54.4% and 63.2% reduction in the total number of biopsies performed compared to the

protocol that every man with suspicion of PCa is biopsied, while the loss in 10-year survival was 0.9% and 1.4%, respectively. Multi-way sensitivity analyses varying base case patients age and serum PSA level in addition to varying the model parameters demonstrated that incorporating PCA3 or T2:ERG into repeat biopsy decisions provided a large reduction in the total number of biopsies (53.2% and 62.0% with a PCA3 threshold of ≥ 25 and T2:ERG threshold of ≥ 10, respectively) and resulted in a small change ($< 2.1\%$) in 10-year overall survival compared to the case where every man was biopsied. The reduction in number of biopsies increased as the threshold for biomarkers increased, while the loss in 10-year survival also increased slightly. In the analysis of 15-year cancer-specific survival, multi-way sensitivity analysis for the base case patient showed that the protocols using a PCA3 threshold of ≥ 25 and a T2:ERG threshold of ≥ 10 at repeat biopsy decisions resulted in the same amount of reduction in the total number of biopsies performed compared to the protocol that every man with suspicion of PCa is biopsied, while the loss in 15-year cancer-specific survival was 1.5% and 2.5%, respectively. Similar to 10-year overall survival analysis, the reduction in number of biopsies increased as the threshold for biomarkers increased, while the loss in 15-year cancer-specific survival also increased slightly.

In order to gain insights about the cost implications of the protocols incorporating biomarkers into repeat biopsy decisions, we calculated the expected cost of the biomarker arm. In the calculations, we assumed the cost of a biopsy and the cost of screening with PSA to be at \$904 [101] and \$30.92 [54], respectively. With the assumption that the institutional costs of PCA3 and T2:ERG markers are the same, we used the bundle cost of \$749 for Mi-Prostate Score (MiPS), which is an early detection test for PCa developed by the researchers in University of Michigan and combines the amount of PSA with the amounts of PCA3 and T2:ERG [168], to estimate the cost of each individual urinary markers. The expected cost of protocols including PCA3 with a threshold of 25 and T2:ERG with a threshold of 10 are estimated to be \$782 and \$753, respectively.

In summary, early detection of PCa must strike a careful balance. The serum PSA test is an existing noninvasive biomarker test that is in common use; however, less predictive than ideal. Although biopsies are much more predictive, they are stressful and traumatic, and can lead to serious complications. The availability of many new biomarker tests has created the opportunity to design new multi-biomarker testing protocols for early diagnosis of PCa; however, the high cost and imperfect sensitivity and specificity of these biomarker tests has raised questions about the most efficient and effective ways to use them.

The work presented in this chapter investigated, for the first time, the value of PCA3

and T2:ERG in the diagnosis of PCa at repeat biopsy by comparing the loss in the overall survival to the gain in repeat biopsy rate. We extend the work of the previous chapters by developing models for the use of multiple continuous biomarkers with the need to identify decision thresholds that account for false positive and false negative results of biomarker tests. The results from this study suggest that PSA alone is ineffective for recommending patients have a repeat biopsy after previous negative biopsies. The addition of PCA3 or T2:ERG for repeat biopsy decisions can reduce the number of biopsies substantially without significantly affecting the patients' survival.

This study has some limitations. We examined a relatively small proportion of patients with clinically insignificant PCa. This raises the question whether our study consists of a representative cohort. However, we need to emphasize that the data was prospectively collected from multiple centers, thus selection-bias is minimal. Additional limitations of this study are related to model inputs such as 10-year overall and 15-year cancer-specific survival, and biopsy sensitivity. The limitations of the studies providing estimates of overall survival include the nonrandomized treatment assignment and retrospective design. The studies evaluated overall survival within 10 years of treatment. A cohort of patients with longer follow-up (more than 10 years) would provide more accurate estimates of long term outcomes. We assumed that the overall survival is independent of the biomarker test scores since there is no study that provides survival estimates considering PCA3 and T2:ERG test results.

These limitations notwithstanding, our study has several strengths, as well as important clinical and policy implications regarding the application of PCA3 assay and T2:ERG in repeat biopsy decisions. We performed a head-to-head comparison of these biomarkers in providing supplementary information to guide repeat biopsy decisions, and found that the PCA3 assay and T2:ERG appear to provide an incremental improvement in the ability to increase the specificity while resulting in a slight decrease in the overall 10-year survival relative to the case where every men is biopsied regardless of the clinical parameters. In addition to the effect on healthcare usage, avoiding unnecessary repeat biopsies will reduce the discomfort, pain, and the other complications associated with repeat biopsies.

Chapter 5.

Summary and Conclusions

The overall objective of this thesis was to leverage the information from new biomarkers and imaging tests to improve early detection of prostate cancer (PCa). Due to the high cost and imperfect predictive accuracy of the tests, how to best use these multiple sources of information is a challenging engineering problem. To address this problem, we developed new optimization-based models and data-driven methods in two primary application areas: (a) development of reliable risk prediction models to provide guidelines for clinical decision-making and the personalized management of PCa, and (b) optimal design of composite diagnostic tests that can account for individualized patient characteristics and uncertainty in disease outcomes using the predictive models. Our models and methods were applied and tested in the context of PCa. However, the utility of these approaches extends to many diseases and other application areas. Following is a summary of the most important findings from Chapters 2, 3 and 4.

In Chapter 2, we combined predictive analytic tools and optimization methods to enable urologists in Michigan to assess individual cancer risk using known clinical risk factors, and to provide clinical recommendations that weigh the benefits and harms of radiological imaging of men with newly-diagnosed PCa. In addition to the published imaging guidelines, we implemented advanced classification modeling techniques to develop accurate classification rules identifying which patients should receive imaging on the basis of individual risk factors. We proposed a new classification algorithm that extracts the information of patients with nonverified disease and incorporates the high cost of misclassifying a metastatic patient simultaneously in its learning framework. In our search for accurate classification rules, we also tested and implemented alternative statistical models that were adapted to improve the classification of imbalanced data through cost-sensitive learning

and resampling techniques. Because not all men with newly-diagnosed PCa underwent a staging bone scan (BS) and computed tomography (CT scan) at diagnosis, we used an established method to correct for the verification bias to evaluate the accuracy of imaging guidelines.

We employed a bi-criteria based approach to determine the Pareto optimal imaging guidelines with respect to the expected number of positive outcomes missed and the expected number of negative tests. We found that the proposed classification model performed well compared to the other classification models that we considered in this work; however, there was no single classification modeling technique that was sufficient with respect to the Pareto optimality criteria. Moreover, the published imaging guidelines were near-optimal for both BS and CT scan. Our work resulted in imaging guidelines for PCa staging that since 2014 are being used across the state of Michigan. We concluded Chapter 2 by describing the post-implementation effects of the proposed guidelines, to confirm their impact on reducing unnecessary imaging.

In Chapter 3, we developed a new sequential testing framework in which some patients may benefit from having tests one at a time so that the results of one test can be used to predict the outcome of the follow-on test. Building on the knowledge attained in Chapter 2, we combined predictive modeling techniques and optimization methods to design coordinated imaging guidelines in which the result of one test may inform the decision about the next test. To incorporate the perspectives of a host of stakeholders (patients, physicians and population) involved in the decision making process for imaging, the goal of the models we developed was to reduce the burden of imaging, while also ensuring that the average risk of missing a metastatic case in the population does not exceed a desirable threshold (the missed-rate budget).

To account for the imperfect calibration of probability estimates obtained from a predictive model, we formulated the decision problem of determining the optimal assignment of patients to imaging protocols as a robust mixed integer program (MIP). We adopted polyhedral uncertainty sets for the model parameters affected by the statistical estimation error, and derived important structural properties of the proposed models. Furthermore, we developed fast, easy-to-understand and clinically motivated heuristics that can mitigate the effects of statistical error by incorporating the knowledge about the estimated risk of metastatic disease. We illustrated the practical performance of the proposed heuristics and optimization models based on medical data collected at a large state-wide collaborative.

Several conclusions with important insights for clinical decision making of PCa imaging

were inferred based on the real case studies. One of the findings was that a greedy algorithm was effective in that it generated more clinically predictable decisions and provided more protection against the missed-rate budget violation compared to optimal solutions to the deterministic model that ignores parameter uncertainty. These results suggested that a simple approximation can achieve clinically acceptable and naturally robust imaging guidelines for the population. Another important finding from this work was that the use of robust coordinated imaging protocols was more beneficial than independently optimized single imaging protocols: they reduced the number of imaging tests performed in the population while providing a significant risk reduction in the missed-rate budget violation. Finally, we showed that there is a significant gap between solutions obtained from the population and individual patient perspectives, suggesting a need for further understanding of ways to bridge this gap.

In contrast to Chapters 2 and 3, in Chapter 4 we focused on diagnostic tests used for early detection of PCa that provide a continuous outcome, rather than a binary (+ or −) outcome, and that are conducted sequentially. We utilized a decision analysis framework and Monte Carlo sampling to determine whether and how to use the newly discovered diagnostic biomarker tests effectively to better select men for repeat biopsy. We used the decision model to examine alternative choices of testing protocols based on the biomarkers with varying thresholds for when to perform additional biomarker tests. Our results suggested that new biomarkers, when used in conjunction with the prostate-specific antigen (PSA) test (with PSA triggering a second a biomarker test), have the potential to reduce the number of biopsies substantially without adversely affecting the overall survival rate of patients with a history of prior negative biopsies. The sensitivity analyses suggested that our conclusions were robust with respect to the plausible variation in the model parameters.

There are several extensions to our work presented in Chapter 2. An important direction is to test the proposed classification model, cost-sensitive Laplacian kernel logistic regression (Cos-LapKLR), for handling unlabeled and imbalanced data simultaneously in other disease areas. The Cos-LapKLR model did not provide significant performance improvement in our context; however, in clinical applications where labeled data is very limited compared to unlabeled data and the accuracy on minority class is the major concern, the Cos-LapKLR model has the potential to improve the detection of patients with high risk of disease. Moreover, the multi-step approach presented in Chapter 2 can used to develop clinical recommendations regarding the use of diagnostic tests in other diseases, that can optimally balance the competing goals of an accurate detection of disease and harms of

testing.

There are several promising variations and extensions to our work presented in Chapter 3. One important direction is to, in addition to the statistical estimation error, incorporate what we call the *model error* into the modeling framework for robust coordinated imaging. The model error results from the use of clinical predictors that are associated with metastatic cancer risk to classify patients into groups for which the proportion of patients in each group must be estimated based on sample data. While in theory this error can be eliminated, in practice it exists because of the strong preference among physicians to have easy-to-understand classifications of patients. Thus, it would be interesting to investigate the impact of model error on the optimal design of coordinated imaging. Furthermore, we focused on patients' health, but not specifically on healthcare costs. Given the serious economic implications of imaging on the patients and on the healthcare system, it is important to understand how the heuristics and optimization models, and the resulting imaging guidelines, change when the costs of imaging tests are taken into account. Results based on the implementation of our findings in Chapter 2 suggest that our models may lead to significant cost savings [82].

With the technological advances in imaging, the spectrum of available imaging options for the management of PCa is continuously evolving. Novel imaging methods such as magnetic resonance imaging (MRI) and combined positron emission tomography (PET)/CT scan have introduced additional options for PCa staging that offer improved sensitivity and specificity, and thus the potential for more accurate assessment of metastatic disease [63, 79, 118]. However, these more advanced techniques are currently being incorporated into clinical use and are not yet widely available. In our work presented in Chapter 3, we studied the optimal design of coordinated imaging guidelines by considering the most commonly used imaging tests (BS and CT scan) for PCa staging. Our results showed that coordinated imaging improves the clinical decision process by achieving better health outcomes and by reducing imaging in the patient population. An important direction would be to extend our modeling framework to incorporate imaging tests that are newly introduced into practice, and to provide insights into their clinical value to improve detection of metastatic cancer.

There are also a number of important avenues of investigation related to our work presented in Chapter 4. An important direction is to extend the decision model to design one-time composite tests consisting of a sequence of biomarker tests. Standard clinical practice assumes simultaneous application of biomarker tests; however, biomarker tests vary in the outcome they predict (all cancer *v.s.* high-grade cancer), in their sensitivities and speci-

ficities, and also vary significantly in cost. Another important source of uncertainty to be included in the model is the imperfect sensitivity of prostate biopsies. In our work, we considered composite tests that combine PSA with an additional biomarker (prostate cancer antigen 3 assay (PCA3) and TMPRSS2:ERG assay (T2:ERG)) and showed that these composite testing strategies provide clinical benefits to patients. Thus, this work provides a starting point for estimating Pareto optimal composite tests that combine PSA with multiple biomarker tests.

In conclusion, we investigated the optimal design of diagnostic testing strategies to determine efficient and effective ways to use individual diagnostic resources. We presented new analytic modeling and algorithmic approaches to achieve an optimal trade-off between the benefits of early detection and the cost and harms of testing, such as unnecessary biopsies. We further tested these models and approaches in the context of PCa to evaluate their potential impact. Moreover, our work provides important insights into how transformative impact can be attained in clinical practice, by addressing the perspectives of multiple stakeholders with varying criteria, including cost (e.g., payers and patients) and clinical criteria (e.g., patients and physicians). The rapid introduction of new discoveries and technologies into routine medical practice has the potential to improve cancer care; however, the enormity of medical data continue to present challenges to improve care delivery and to reduce wasteful utilization of diagnostic resources. The work presented in this thesis could help lay the groundwork to improve early detection of other types of cancer and other diseases by leveraging information from multiple sources of testing.

Appendix A.

Supplements to Chapter 2

A.1. Results for Random Forests and Adaboost

Several data balancing techniques exist in literature to deal with the class imbalance problem in different forms of resampling. Two non-heuristic sampling methods are commonly used: random oversampling (ROS) of the minority class and random undersampling (RUS) of the majority class.

The Synthetic Minority Oversampling Technique (SMOTE) is a method of oversampling, which produces synthetic minority instances by selecting some of the nearest minority neighbors of a minority instance and generating synthetic minority instance along with the lines between the minority instance and the nearest minority neighbors [39]. Although it has shown many promising benefits, the SMOTE algorithm also has drawbacks, such as overfitting. It introduces the same number of synthetic patients for each minority patient without considering the neighboring patients, which increases the occurrence of overlapping between minority and majority class. Borderline-SMOTE was proposed to enhance the original concept by identifying the borderline minority samples [70]. In order to obtain well-defined class clusters, several data cleaning methods such as the Edited Nearest Neighbor (ENN) rule [12] and Tomek links [166] have been integrated with SMOTE. SMOTE combined with two data cleaning techniques, Tomek links and ENN Rule [176], have shown better performance in data sets with a small number of minority instances.

To improve upon the performance of random undersampling, several undersampling methods combined with data cleaning techniques have been proposed such as Tomek links, Condensed Nearest Neighbor Rule (CNN) [72] and Neighborhood Cleaning Rule (NCR) [91]. In this work, we implement and test ten different methods of under and oversampling

to balance the class distribution on training data.

These methods are available in the `imbalanced-learn` package in Python [93]. We used this package to perform 10 independent runs of 2-fold cross validation (CV) on the development samples. The results from these experiments are summarized in Table A.1. The experimental results indicate that the accuracy of classification rules on the bone scan (BS) and computed tomography (CT scan) data sets developed by Random forests (RF) and AdaBoost can be improved via model-independent data-driven approaches. For instance, the baseline RF identifying patients with bone metastasis obtained a sensitivity of 24.97% and specificity of 98.05%, whereas RF combined with RUS improved the sensitivity to 74.68% while reducing the specificity to 68.13%. RF and Adaboost combined with RUS achieved the highest sensitivity and area under the ROC curve (AUC) in both BS and CT scan datasets. These results clearly illustrate the inadequacy of the baseline Random forests and AdaBoost in recognizing metastatic patients.

Table A.1: Performance of RF and AdaBoost for BS and CT scan in 10 independent repetitions of 2-fold CV.

	Bone scan (n = 416)				**CT (n = 643)**			
Models	**Sensitivity**	**Specificity**	**AUC**	**Brier**	**Sensitivity**	**Specificity**	**AUC**	**Brier**
RF								
Original	24.97	98.05	79.35	0.087	32.68	98.18	86.80	0.062
RUS	74.68	68.13	78.88	0.20	75.19	77.22	84.20	0.16
CNN	34.68	94.44	76.53	0.11	45.36	96.54	86.51	0.076
NCR	40.95	93.47	79.47	0.096	46.44	95.72	85.79	0.070
Tomek Links	28.54	97.19	79.92	0.086	38.65	97.71	86.55	0.062
ROS	32.46	94.53	77.44	0.099	36.94	96.70	85.62	0.069
SMOTE	41.83	89.35	78.32	0.12	40.37	94.64	84.68	0.080
SMOTE-Borderline	44.10	90.78	78.44	0.11	40.07	95.16	85.06	0.078
SMOTE + Tomek links	45.11	88.80	78.16	0.12	40.63	94.47	84.83	0.080
SMOTE + ENN	65.56	78.16	79.37	0.17	56.80	83.52	82.89	0.14
AdaBoost								
Original	18.78	95.63	64.29	0.24	33.91	96.55	80.87	0.24
RUS	62.67	62.13	68.87	0.24	71.64	73.10	81.08	0.22
CNN	33.41	84.85	61.86	0.24	43.99	84.69	75.21	0.24
NCR	38.62	92.42	76.37	0.23	43.63	95.69	80.74	0.23
Tomek Links	28.31	95.66	71.34	0.24	38.45	96.55	80.87	0.24
ROS	19.15	95.01	64.79	0.24	38.77	95.03	80.44	0.24
SMOTE	32.51	88.72	63.71	0.24	45.25	92.16	79.17	0.24
SMOTE-Borderline	35.13	89.91	66.29	0.24	42.08	92.40	79.53	0.24
SMOTE + Tomek links	33.84	87.63	64.76	0.24	43.23	91.58	78.64	0.24
SMOTE + ENN	65.98	74.90	79.14	0.23	63.44	83.98	81.99	0.23

Sensitivity, specificity and AUC are reported in percentages.

Appendix B.

Supplements to Chapter 3

B.1. Results for Multinomial Model

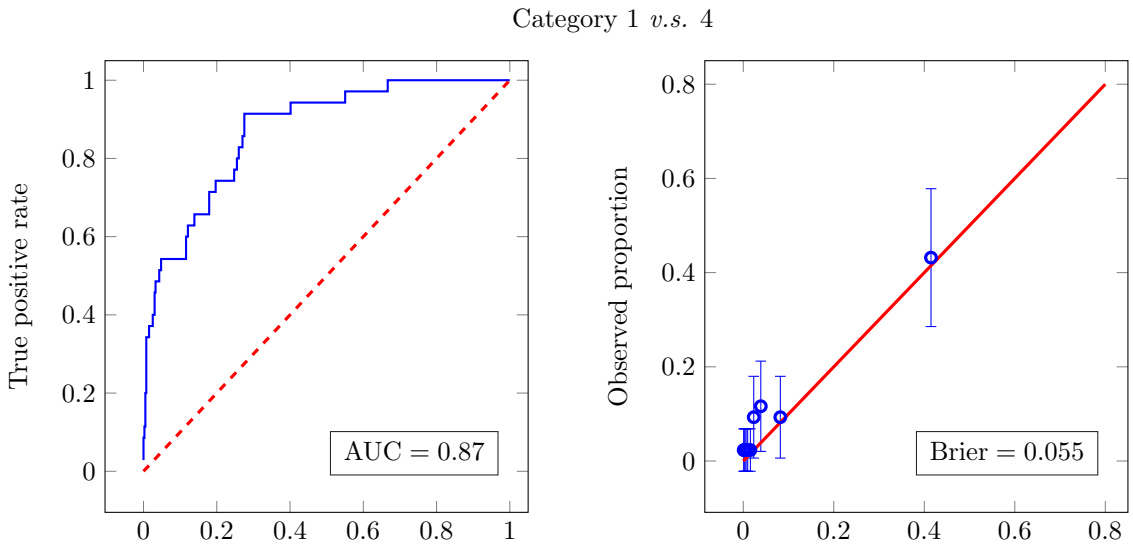

Figure B.1: Pairwise performance comparisons of binary models based on the validation samples for category 1 *v.s.* 4.

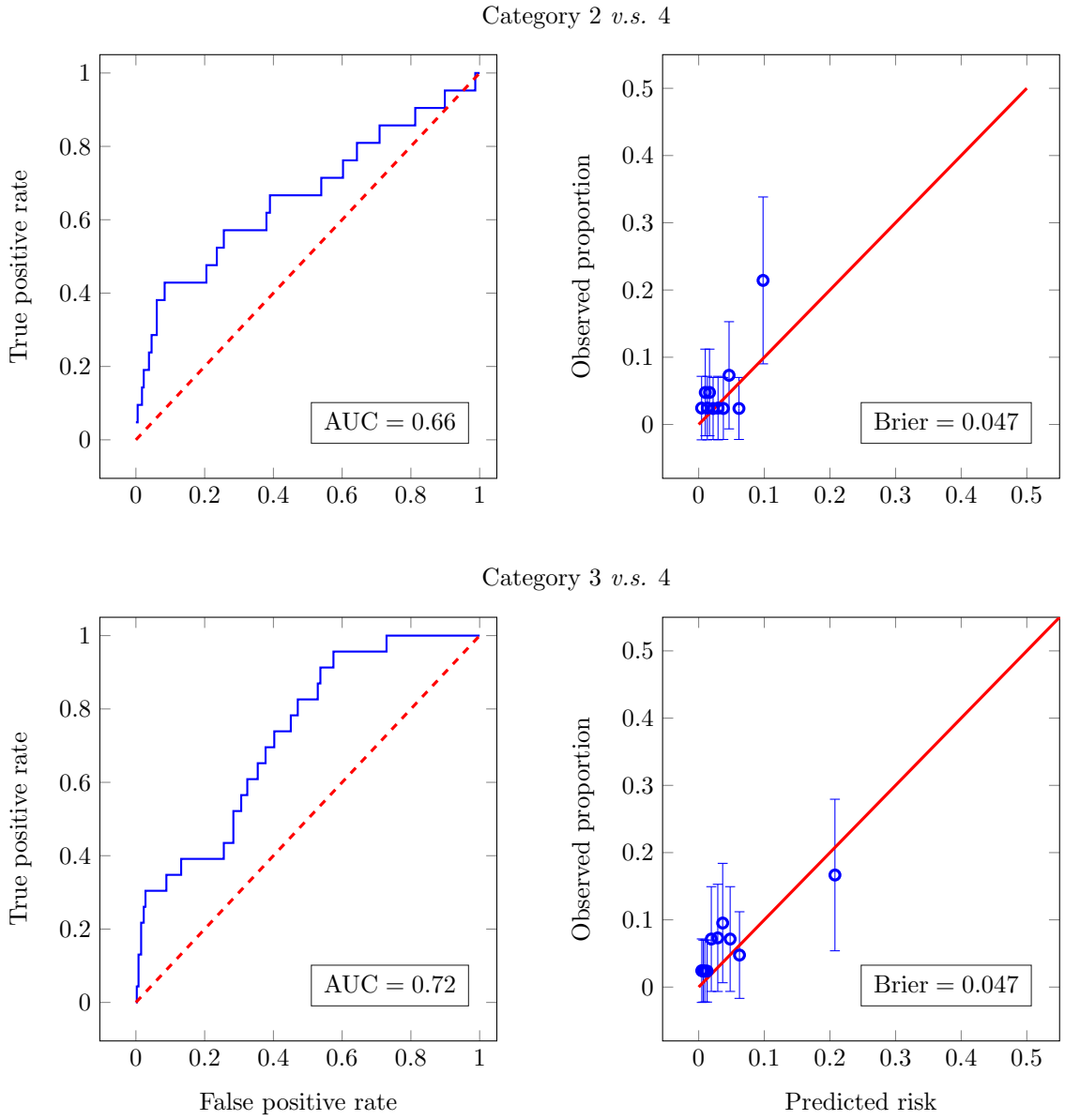

Figure B.1: Pairwise performance comparisons of binary models based on the validation samples for category 2 *v.s.* 4 and category 3 *v.s.* 4.

B.2. Results for Optimization Models

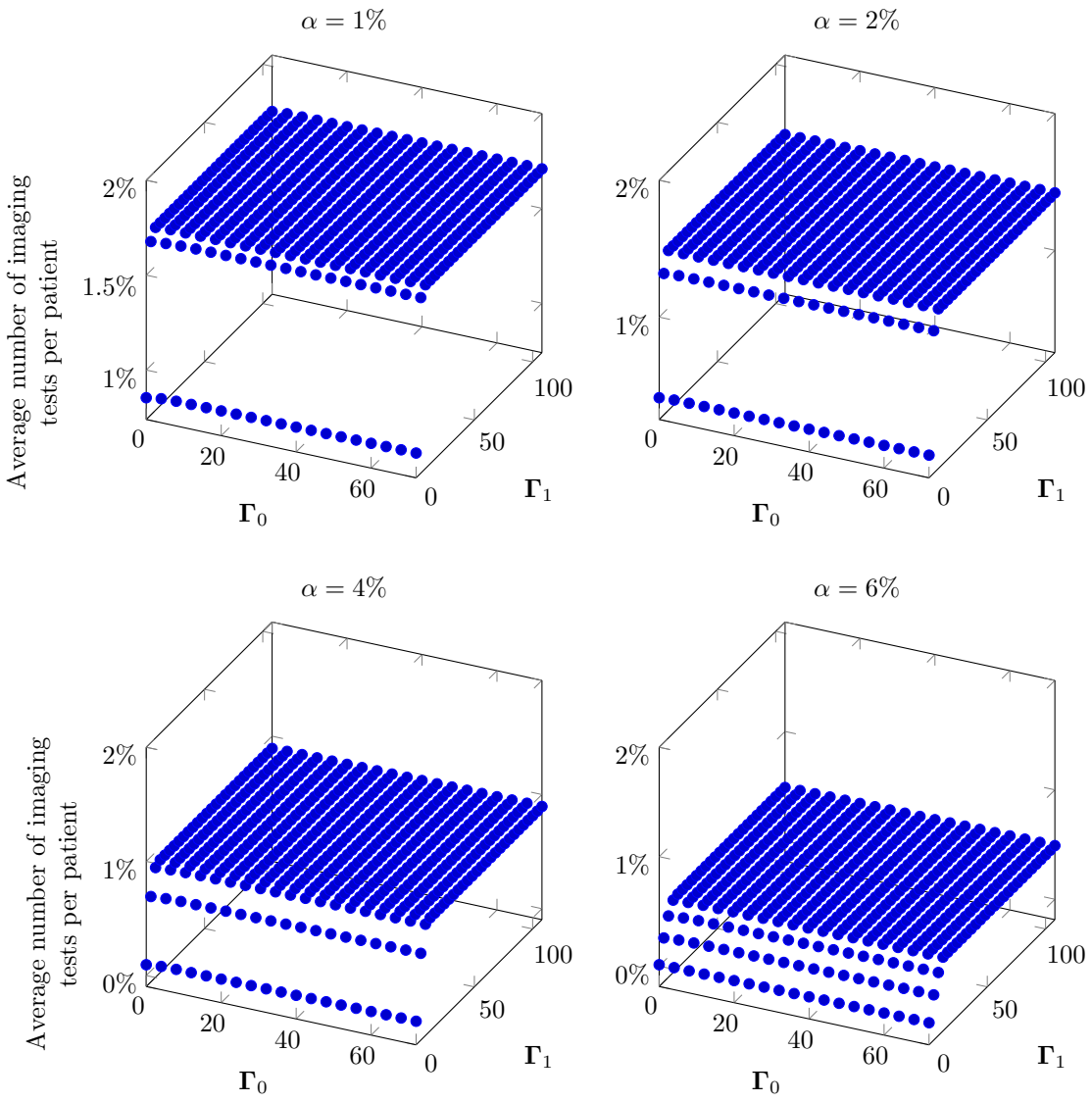

Figure B.2: The effect of the protection levels $\mathbf{\Gamma}_0$ and $\mathbf{\Gamma}_1$ on the optimality of solutions to R-MIM.

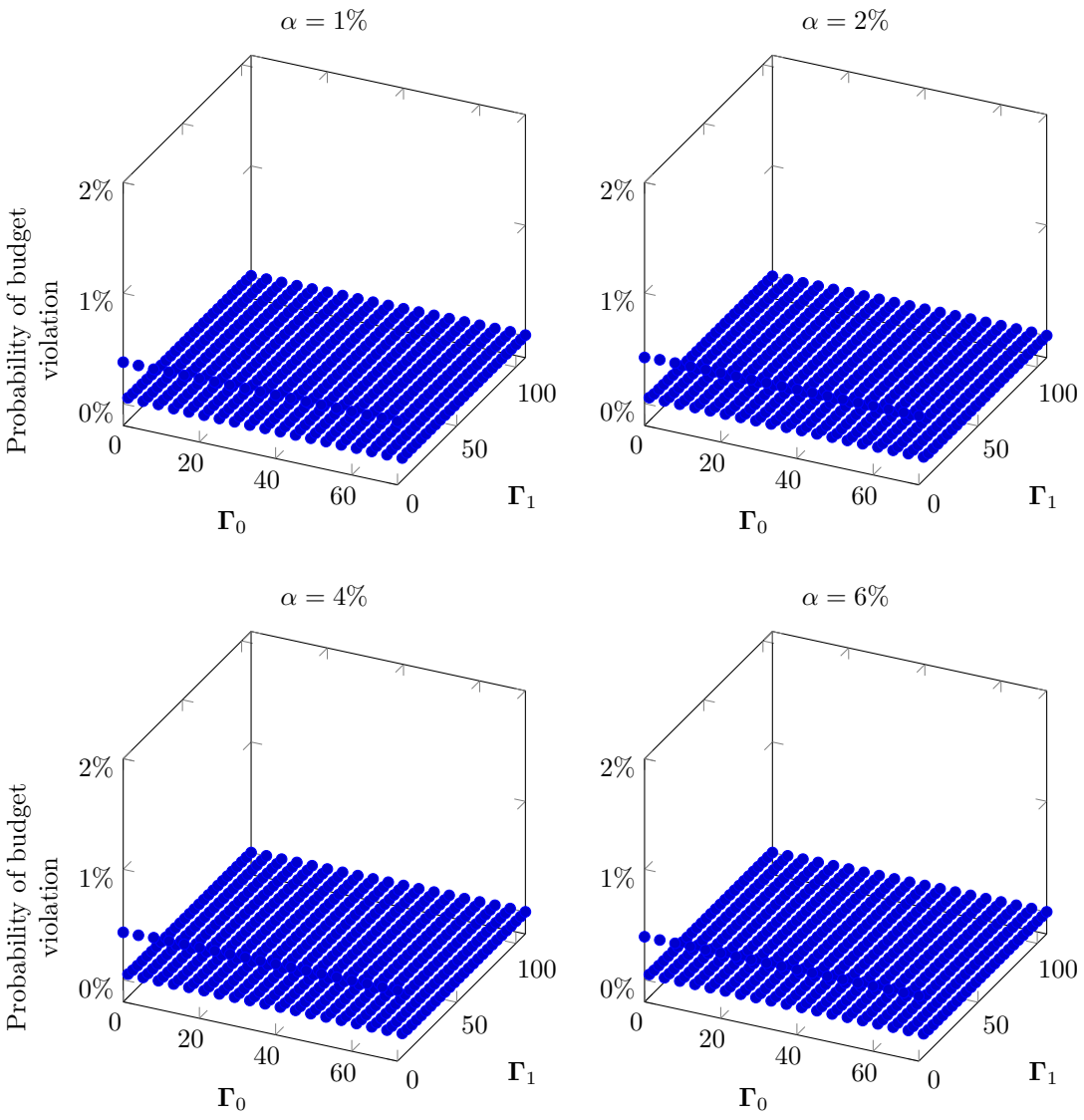

Figure B.3: The effect of the protection levels $\mathbf{\Gamma}_0$ and $\mathbf{\Gamma}_1$ on the robustness of solutions to R-MIM.

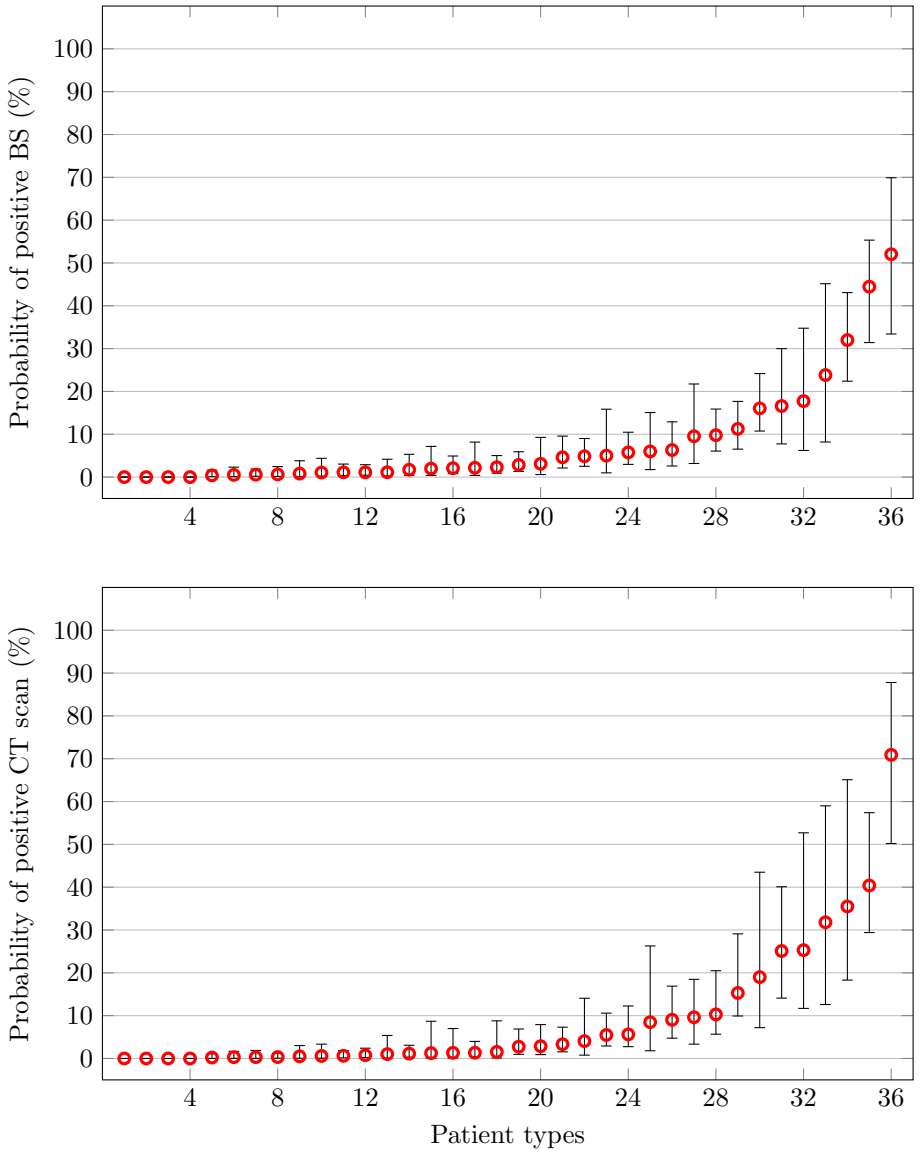

Figure B.4: Illustration of the ranges in the estimated probability of positive bone scan (BS) and computed tomography (CT scan) for patient types. The patient types are sorted in the order of increasing risk of disease.

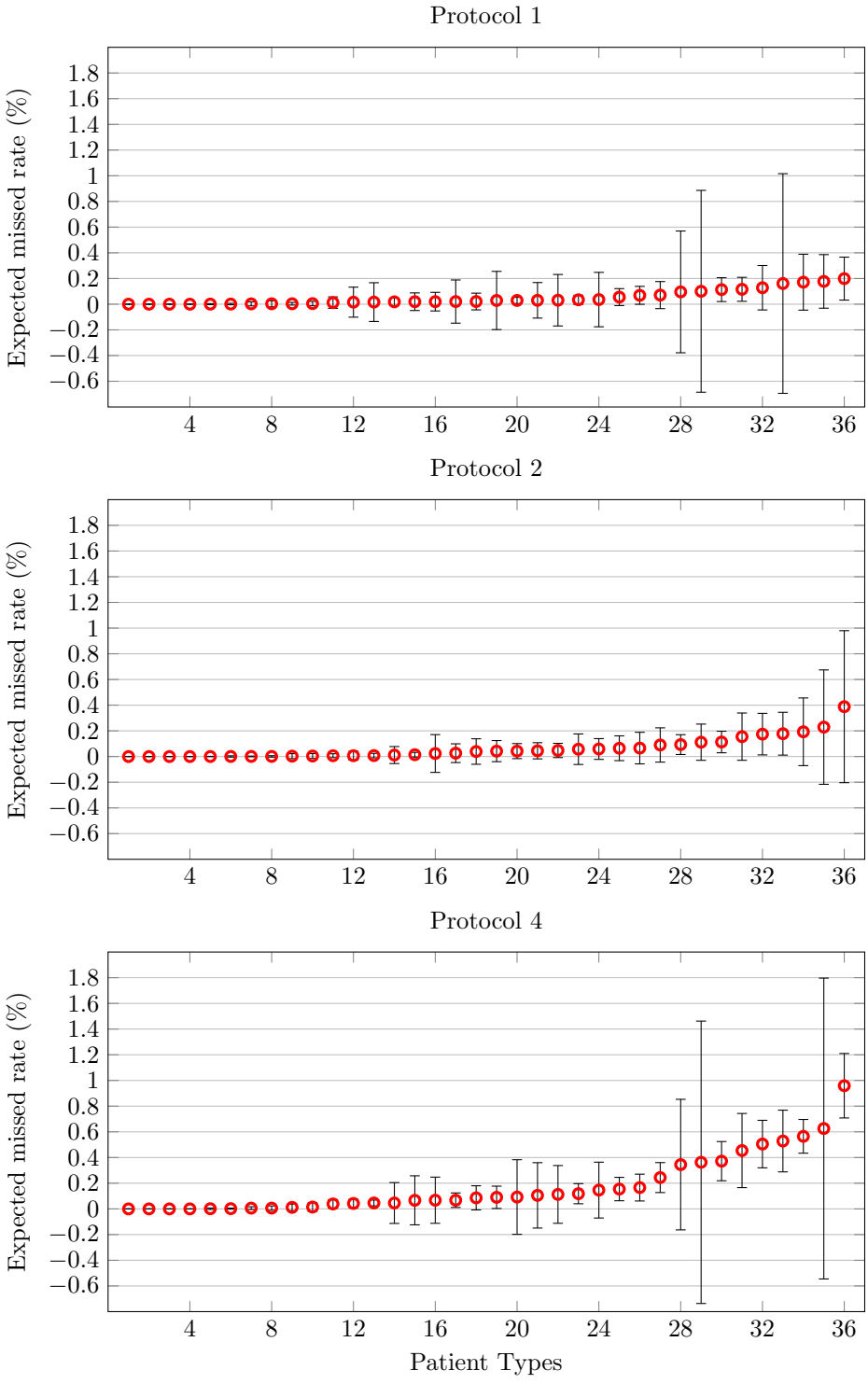

Figure B.5: Expected missed disease rates for Protocols $1, 2$ and 4. The patient types are sorted in the order of increasing expected missed disease rate in each panel.

Appendix C.

Supplements to Chapter 4

C.1. Sensitivity Analyses

Table C.1: Assumed distributions and parameters used in probabilistic sensitivity analysis.

	Point Estimate (μ)	Range	Standard deviation (σ)	α	β
Probabilities*					
Patient has GS < 7 cancer given that he actually has prostate cancer	0.59	$[0.29 - 1]$	0.042	81.41	57.59
Patient has GS $= 7$ cancer given that he actually has prostate cancer	0.30	$[0.15 - 0.60]$	0.039	41.70	97.30
Patient has GS > 7 cancer given that he actually has prostate cancer	0.11	$[0.06 - 0.23]$	0.027	15.89	123.11
Patient has PCA3 score ≥ 25 given that he actually has GS < 7 cancer	0.65	$[0.32 - 1]$	0.053	52.35	28.65
Patient has PCA3 score ≥ 25 given that he actually has GS $= 7$ cancer	0.74	$[0.37 - 1]$	0.068	30.26	10.74
Patient has PCA3 score ≥ 25 given that he actually has GS > 7 cancer	0.69	$[0.34 - 1]$	0.116	10.31	4.69
Patient has PCA3 score ≥ 35 given that he actually has GS < 7 cancer	0.45	$[0.23 - 0.90]$	0.055	36.55	44.45
Patient has PCA3 score ≥ 35 given that he actually has GS $= 7$ cancer	0.55	$[0.27 - 1]$	0.077	22.45	18.55
Patient has PCA3 score ≥ 35 given that he actually has GS > 7 cancer	0.56	$[0.28 - 1]$	0.124	8.44	6.56
Patient has T2:ERG ≥ 7 given that he actually has GS < 7 cancer	0.52	$[0.26 - 1]$	0.055	42.48	38.52
Patient has T2:ERG ≥ 7 given that he actually has GS $= 7$ cancer	0.55	$[0.27 - 1]$	0.077	22.45	18.55
Patient has T2:ERG ≥ 7 given that he actually has GS > 7 cancer	0.75	$[0.38 - 1]$	0.108	11.25	3.75
Patient has T2:ERG ≥ 10 given that he actually has GS < 7 cancer	0.46	$[0.23 - 0.93]$	0.055	37.54	43.46
Patient has T2:ERG ≥ 10 given that he actually has GS $= 7$ cancer	0.50	$[0.25 - 1]$	0.077	20.50	20.50
Patient has T2:ERG ≥ 10 given that he actually has GS > 7 cancer	0.75	$[0.38 - 1]$	0.108	11.25	3.75
Patient has T2:ERG ≥ 30 given that he actually has GS < 7 cancer	0.37	$[0.18 - 0.73]$	0.053	29.63	51.37
Patient has T2:ERG ≥ 30 given that he actually has GS $= 7$ cancer	0.29	$[0.14 - 0.57]$	0.070	11.71	29.29
Patient has T2:ERG ≥ 30 given that he actually has GS > 7 cancer	0.50	$[0.25 - 1]$	0.125	7.50	7.50
Patient has T2:ERG ≥ 50 given that he actually has GS < 7 cancer	0.26	$[0.13 - 0.51]$	0.048	20.74	60.26
Patient has T2:ERG ≥ 50 given that he actually has GS $= 7$ cancer	0.26	$[0.13 - 0.52]$	0.068	10.74	30.26
Patient has T2:ERG ≥ 50 given that he actually has GS > 7 cancer	0.50	$[0.25 - 1]$	0.125	7.50	7.50
Patient has T2:ERG ≥ 100 given that he actually has GS < 7 cancer	0.16	$[0.08 - 0.32]$	0.040	12.84	68.16
Patient has T2:ERG ≥ 100 given that he actually has GS $= 7$ cancer	0.17	$[0.08 - 0.33]$	0.058	6.83	34.17
Patient has T2:ERG ≥ 100 given that he actually has GS > 7 cancer	0.38	$[0.19 - 0.75]$	0.121	5.63	9.38
Patient has PCA3 score ≥ 25 given that he does not have cancer	0.38	$[0.19 - 0.75]$	0.121	5.63	9.38
Patient has PCA3 score ≥ 35 given that he does not have cancer	0.26	$[0.13 - 0.52]$	0.109	3.86	11.14
Patient has T2:ERG ≥ 7 given that he does not have cancer	0.38	$[0.19 - 0.76]$	0.121	5.68	9.32
Patient has T2:ERG ≥ 10 given that he does not have cancer	0.31	$[0.16 - 0.63]$	0.116	4.71	10.29
Patient has T2:ERG ≥ 30 given that he does not have cancer	0.20	$[0.10 - 0.40]$	0.100	3.00	12.00
Patient has T2:ERG ≥ 50 given that he does not have cancer	0.16	$[0 - 0.33]$	0.093	2.46	12.54
Patient has T2:ERG ≥ 100 given that he does not have cancer	0.08	$[0 - 0.15]$	0.066	1.13	13.88
Biopsy Sensitivity	0.80	$[0.40 - 1]$	0.090	15.20	3.80
Patient parameters†					
Serum PSA (ng/ml)	–	$[4 - 30]$	5.78	2.55	2.96
Age (years)	–	$[50 - 85]$	8.45	60.0	1.10

*All the assumed distributions for probability estimates are beta distributions; †The assumed distributions for patient parameters are gamma distributions.

Table C.2: Assumed distributions and parameters used in probabilistic sensitivity analysis.

Overall 10-year survival*	Point Estimate (μ)	Range	Standard deviation (σ)	α	β
Age: Younger Than 60 years					
Survival without PCa	0.94	$[0.47 - 1]$	0.005	2035.72	129.94
PSA: $0.0 - 9.9$					
Untreated survival for GS $<$ 7 cancer	0.85	$[0.43 - 1]$	0.04	84.12	14.84
Untreated survival for GS $=$ 7 cancer	0.81	$[0.41 - 1]$	0.04	96.92	22.73
Untreated survival for GS $>$ 7 cancer	0.77	$[0.39 - 1]$	0.05	51.62	15.42
Treated survival for GS $<$ 7 cancer	0.94	$[0.47 - 1]$	0.02	225.36	14.38
Treated survival for GS $=$ 7 cancer	0.92	$[0.46 - 1]$	0.02	288.10	25.05
Treated survival for GS $>$ 7 cancer	0.90	$[0.45 - 1]$	0.03	123.57	13.73
PSA: $10 - 19.9$					
Untreated survival for GS $<$ 7 cancer	0.76	$[0.38 - 1]$	0.05	52.49	16.58
Untreated survival for GS $=$ 7 cancer	0.71	$[0.36 - 1]$	0.05	45.70	18.67
Untreated survival for GS $>$ 7 cancer	0.65	$[0.33 - 1]$	0.07	28.33	15.26
Treated survival for GS $<$ 7 cancer	0.90	$[0.45 - 1]$	0.03	123.57	13.73
Treated survival for GS $=$ 7 cancer	0.87	$[0.44 - 1]$	0.03	150.33	22.46
Treated survival for GS $>$ 7 cancer	0.84	$[0.42 - 1]$	0.04	87.67	16.70
PSA: 20 or Greater					
Untreated survival for GS $<$ 7 cancer	0.73	$[0.37 - 1]$	0.06	37.65	13.93
Untreated survival for GS $=$ 7 cancer	0.67	$[0.34 - 1]$	0.06	33.00	16.26
Untreated survival for GS $>$ 7 cancer	0.60	$[0.30 - 1]$	0.07	27.62	18.42
Treated survival for GS $<$ 7 cancer	0.88	$[0.44 - 1]$	0.03	98.28	13.40
Treated survival for GS $=$ 7 cancer	0.85	$[0.43 - 1]$	0.04	84.12	14.84
Treated survival for GS $>$ 7 cancer	0.81	$[0.41 - 1]$	0.04	74.02	17.36
Age: 61 - 70 years					
Survival without PCa	0.85	$[0.43 - 1]$	0.01	1039.98	183.53
PSA: $0.0 - 9.9$					
Untreated survival for GS $<$ 7 cancer	0.75	$[0.38 - 1]$	0.03	149.31	49.77
Untreated survival for GS $=$ 7 cancer	0.70	$[0.35 - 1]$	0.04	87.54	37.52
Untreated survival for GS $>$ 7 cancer	0.63	$[0.32 - 1]$	0.05	55.79	32.76
Treated survival for GS $<$ 7 cancer	0.89	$[0.45 - 1]$	0.02	371.02	45.86
Treated survival for GS $=$ 7 cancer	0.87	$[0.44 - 1]$	0.02	235.38	35.17
Treated survival for GS $>$ 7 cancer	0.83	$[0.42 - 1]$	0.03	179.13	36.69
PSA: $10 - 19.9$					
Untreated survival for GS $<$ 7 cancer	0.62	$[0.31 - 1]$	0.05	68.66	42.08
Untreated survival for GS $=$ 7 cancer	0.55	$[0.28 - 1]$	0.05	51.74	42.34
Untreated survival for GS $>$ 7 cancer	0.47	$[0.24 - 1]$	0.06	30.76	34.69
Treated survival for GS $<$ 7 cancer	0.83	$[0.42 - 1]$	0.03	124.14	25.43
Treated survival for GS $=$ 7 cancer	0.79	$[0.40 - 1]$	0.03	139.07	36.97
Treated survival for GS $>$ 7 cancer	0.74	$[0.37 - 1]$	0.04	84.72	29.77
PSA: 20 or Greater					
Untreated survival for GS $<$ 7 cancer	0.57	$[0.29 - 1]$	0.06	43.79	33.03
Untreated survival for GS $=$ 7 cancer	0.49	$[0.25 - 1]$	0.05	46.55	48.45
Untreated survival for GS $>$ 7 cancer	0.40	$[0.20 - 0.80]$	0.05	36.48	54.72
Treated survival for GS $<$ 7 cancer	0.8	$[0.40 - 1]$	0.04	99.55	24.89
Treated survival for GS $=$ 7 cancer	0.75	$[0.38 - 1]$	0.03	149.31	49.77
Treated survival for GS $>$ 7 cancer	0.70	$[0.35 - 1]$	0.04	87.54	37.52
Age: Older than 70 years					
Survival without PCa	0.77	$[0.39 - 1]$	0.015	581.30	173.64
PSA: $0.0 - 9.9$					
Untreated survival for GS $<$ 7 cancer	0.66	$[0.33 - 1]$	0.05	69.58	35.85
Untreated survival for GS $=$ 7 cancer	0.59	$[0.30 - 1]$	0.05	67.10	46.63
Untreated survival for GS $>$ 7 cancer	0.51	$[0.26 - 1]$	0.06	39.95	38.39
Treated survival for GS $<$ 7 cancer	0.85	$[0.43 - 1]$	0.03	165.68	29.24
Treated survival for GS $=$ 7 cancer	0.81	$[0.41 - 1]$	0.03	132.22	31.01
Treated survival for GS $>$ 7 cancer	0.76	$[0.38 - 1]$	0.05	68.12	26.49
PSA: $10 - 19.9$					
Untreated survival for GS $<$ 7 cancer	0.5	$[0.25 - 1]$	0.06	39.19	39.19
Untreated survival for GS $=$ 7 cancer	0.41	$[0.21 - 0.82]$	0.05	37.69	54.24
Untreated survival for GS $>$ 7 cancer	0.32	$[0.16 - 0.64]$	0.05	26.43	56.16
Treated survival for GS $<$ 7 cancer	0.76	$[0.38 - 1]$	0.04	82.45	26.04
Treated survival for GS $=$ 7 cancer	0.70	$[0.35 - 1]$	0.04	87.54	37.52
Treated survival for GS $>$ 7 cancer	0.64	$[0.32 - 1]$	0.05	69.58	35.85
PSA: 20 or Greater					
Untreated survival for GS $<$ 7 cancer	0.44	$[0.22 - 0.88]$	0.06	28.48	36.25
Untreated survival for GS $=$ 7 cancer	0.35	$[0.18 - 0.70]$	0.05	37.41	69.48
Untreated survival for GS $>$ 7 cancer	0.26	$[0.13 - 0.52]$	0.05	23.46	66.78
Treated survival for GS $<$ 7 cancer	0.72	$[0.36 - 1]$	0.05	68.12	26.49
Treated survival for GS $=$ 7 cancer	0.66	$[0.33 - 1]$	0.05	56.01	31.50
Treated survival for GS $>$ 7 cancer	0.59	$[0.30 - 1]$	0.05	44.72	31.08

*All the assumed distributions for overall 10-year survival estimates are beta distributions.

Table C.3: Assumed distributions and parameters for 15-year cancer-specific survivals used in probabilistic sensitivity analysis.

15-year cancer-specific survival*	**Point Estimate (μ)**	**Range**	**Standard deviation (σ)**	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$
Survival without PCa	0.94	$[0.47 - 1]$	0.01	508.23	32.44
Survival treated without curative intent					
Untreated survival for GS < 7 cancer	0.65	$[0.33 - 1]$	0.09	18.31	10.08
Untreated survival for GS $= 7$ cancer	0.65	$[0.33 - 1]$	0.09	18.31	10.08
Untreated survival for GS > 7 cancer	0.29	$[0.14 - 1]$	0.06	3.96	9.87
Survival after radical prostatectomy					
Treated survival for GS < 7 cancer	0.94	$[0.47 - 1]$	0.01	508.23	32.44
Treated survival for GS $= 7$ cancer	0.83	$[0.42 - 1]$	0.05	54.71	11.21
Treated survival for GS > 7 cancer	0.66	$[0.33 - 1]$	0.06	38.85	20.01

All the assumed distributions for overall 15-year survival estimates are beta distributions.

Bibliography

[1] ACC. *American Cancer Society - Breast Cancer Statistics*. https://www.cancer.org/research/cancer-facts-statistics/breast-cancer-facts-figures.html. 2018.

[2] ACC. *American Cancer Society - Prostate Cancer Statistics*. https://www.cancer.org/cancer/prostate-cancer/about/key-statistics.html. 2018.

[3] O. Akin and H. Hricak. "Imaging of prostate cancer". In: *Radiologic Clinics* 45.1 (2007), pp. 207–222.

[4] P. C. Albertsen, D. F. Moore, W. Shih, Y. Lin, H. Li, and G. L. Lu-Yao. "Impact of comorbidity on survival among men with localized prostate cancer". In: *Journal of Clinical Oncology* 29.10 (2011), pp. 1335–1341.

[5] H. El-Amine, E. K. Bish, and D. R. Bish. "Robust Postdonation Blood Screening Under Prevalence Rate Uncertainty". In: *Operations Research* (2017), pp. 1–17. DOI: 10.1287/opre.2017.1658. eprint: https://doi.org/10.1287/opre.2017.1658. URL: https://doi.org/10.1287/opre.2017.1658.

[6] AUA. *American Urological Association - Choosing Wisely Campaign*. https://www.auanet.org/practice-resources/patient-safety-and-quality-of-care/choosing-wisely. 2017.

[7] S. M. Aubin, J. Reid, M. J. Sarno, A. Blase, J. Aussie, H. Rittenhouse, R. Rittmaster, G. L. Andriole, and J. Groskopf. "PCA3 molecular urine test for predicting repeat prostate biopsy outcome in populations at risk: validation in the placebo arm of the dutasteride REDUCE trial". In: *The Journal of Urology* 184.5 (2010), pp. 1947–1952.

[8] M. Auprich, H. Augustin, L. Budaus, L. Kluth, S. Mannweiler, S. F. Shariat, M. Fisch, M. Graefen, K. Pummer, and F. K. Chun. "A comparative performance analysis of total prostate-specific antigen, percentage free prostate-specific antigen, prostate-specific antigen velocity and urinary prostate cancer gene 3 in the first, second and third repeat prostate biopsy". In: *BJU International* 109.11 (2012), pp. 1627–1635.

[9] M. Auprich, F. K. Chun, J. F. Ward, K. Pummer, R. Babaian, H. Augustin, F. Luger, S. Gutschi, L. Budaus, M. Fisch, et al. "Critical assessment of preoperative urinary prostate cancer antigen 3 on the accuracy of prostate cancer staging". In: *European Urology* 59.1 (2011), pp. 96–105.

[10] M. Auprich, A. Haese, J. Walz, K. Pummer, A. de la Taille, M. Graefen, T. de Reijke, M. Fisch, P. Kil, P. Gontero, et al. "External validation of urinary PCA3-based nomograms to individually predict prostate biopsy outcome". In: *European Urology* 58.5 (2010), pp. 727–732.

[11] M. J. Barry. "Prostate-specific–antigen testing for early diagnosis of prostate cancer". In: *New England Journal of Medicine* 344.18 (2001), pp. 1373–1377.

[12] G. E. Batista, R. C. Prati, and M. C. Monard. "A study of the behavior of several methods for balancing machine learning training data". In: *ACM SIGKDD Explorations Newsletter* 6.1 (2004), pp. 20–29.

[13] C. B. Begg. "Biases in the assessment of diagnostic tests". In: *Statistics in Medicine* 6.4 (1987), pp. 411–423.

[14] C. B. Begg and R. A. Greenes. "Assessment of diagnostic tests when disease verification is subject to selection bias". In: *Biometrics* (1983), pp. 207–215.

[15] M. Belkin, P. Niyogi, and V. Sindhwani. "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples". In: *The Journal of Machine Learning Research* 7 (2006), pp. 2399–2434.

[16] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton University Press, 2009.

[17] A. Ben-Tal and A. Nemirovski. "Robust convex optimization". In: *Mathematics of operations research* 23.4 (1998), pp. 769–805.

[18] A. Ben-Tal and A. Nemirovski. "Robust solutions of linear programming problems contaminated with uncertain data". In: *Mathematical programming* 88.3 (2000), pp. 411–424.

[19] A. Ben-Tal and A. Nemirovski. "Robust solutions of uncertain linear programs". In: *Operations Research Letters* 25.1 (1999), pp. 1–13.

[20] A. Ben-Tal and A. Nemirovski. "Selected topics in robust convex optimization". In: *Mathematical Programming* 112.1 (2008), pp. 125–158.

[21] D. Bertsimas, D. B. Brown, and C. Caramanis. "Theory and Applications of Robust Optimization". In: *SIAM Review* 53.3 (2011), pp. 464–501. DOI: 10.1137/080734510. eprint: http://dx.doi.org/10.1137/080734510. URL: http://dx.doi.org/10.1137/080734510.

[22] D. Bertsimas and M. Sim. "Robust discrete optimization and network flows". In: *Mathematical Programming* 98.1-3 (2003), pp. 49–71. ISSN: 0025-5610. DOI: 10.1007/s10107-003-0396-4. URL: http://dx.doi.org/10.1007/s10107-003-0396-4.

[23] D. Bertsimas and M. Sim. "The Price of Robustness". In: *Operations Research* 52.1 (2004), pp. 35–53. ISSN: 0030364X, 15265463. URL: http://www.jstor.org/stable/30036559.

[24] J. R. Birge and F. Louveaux. *Introduction to stochastic programming*. Springer Science & Business Media, 2011.

[25] D. R. Bish, E. K. Bish, R. S. Xie, and S. L. Stramer. "Going beyond "same-for-all" testing of infectious agents in donated blood". In: *IIE Transactions* 46.11 (2014), pp. 1147–1168.

[26] D. R. Bish, E. K. Bish, S. R. Xie, and A. D. Slonim. "Optimal selection of screening assays for infectious agents in donated blood". In: *IIE Transactions on Healthcare Systems Engineering* 1.2 (2011), pp. 67–90.

[27] S. Bleeker, H. Moll, E. Steyerberg, A. Donders, G. Derksen-Lubsen, D. Grobbee, and K. Moons. "External validation is necessary in prediction research:: A clinical example". In: *Journal of Clinical Epidemiology* 56.9 (2003), pp. 826–832.

[28] M. Blute and D. Jason Efstathiou MD. "Management of lymph node-positive prostate cancer: the role of surgery and radiation therapy". In: *Oncology* 27.7 (2013), p. 647.

[29] L. Breiman. "Random forests". In: *Machine Learning* 45.1 (2001), pp. 5–32.

[30] A. Briganti, N. Passoni, M. Ferrari, U. Capitanio, N. Suardi, A. Gallina, L. F. Da Pozzo, M. Picchio, V. Di Girolamo, A. Salonia, et al. "When to perform bone scan in patients with newly diagnosed prostate cancer: external validation of the currently available guidelines and proposal of a novel risk stratification tool". In: *European Urology* 57.4 (2010), pp. 551–558.

[31] M. Broder, B. Gutierrez, D. Cherepanov, and Y. Linhares. "Burden of skeletal-related events in prostate cancer: unmet need in pain improvement". In: *Supportive Care in Cancer* 23.1 (2015), pp. 237–247.

[32] L. Bubendorf, A. Schöpfer, U. Wagner, G. Sauter, H. Moch, N. Willi, T. C. Gasser, and M. J. Mihatsch. "Metastatic patterns of prostate cancer: an autopsy study of 1,589 patients". In: *Human Pathology* 31.5 (2000), pp. 578–583.

[33] P. S. Bunting. "Screening for prostate cancer with prostatespecific antigen: beware the biases". In: *Clinica Chimica Acta* 315.1 (2002), pp. 71–97.

[34] M. J. Bussemakers, A. van Bokhoven, G. W. Verhaegh, F. P. Smit, H. F. Karthaus, J. A. Schalken, F. M. Debruyne, N. Ru, and W. B. Isaacs. "DD3:: A new prostate-specific gene, highly overexpressed in prostate cancer". In: *Cancer Research* 59.23 (1999), pp. 5975–5979.

[35] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. "A limited memory algorithm for bound constrained optimization". In: *SIAM Journal on Scientific Computing* 16.5 (1995), pp. 1190–1208.

[36] P. Carroll, K. Greene, R. J. Babaian, P. H. G. H. Ballentine Carter, M. Han, D. A. Kuban, A. O. Sartor, J. L. Stanford, and A. Zietman. "PSA Testing for the Pretreatment Staging and Posttreatment Management of Prostate Cancer: 2013 Revision of 2009 Best Practice Statement". In: *American Urological Association* (2013). URL: `https://www.auanet.org/guidelines/prostate-specific-antigen-(2009-amended-2013)`.

[37] W. J. Catalona, J. P. Richie, F. R. Ahmann, A. H. M'Liss, P. T. Scardino, R. C. Flanigan, J. B. Dekernion, T. L. Ratliff, L. R. Kavoussi, B. L. Dalkin, et al. "Comparison of digital rectal examination and serum prostate specific antigen in the early detection of prostate cancer: results of a multicenter clinical trial of 6,630 men". In: *The Journal of Urology* 151.5 (1994), pp. 1283–1290.

[38] O. Chapelle, B. Schlkopf, and A. Zien. *Semi-Supervised Learning*. 1st. The MIT Press, 2010.

[39] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. "SMOTE: synthetic minority over-sampling technique". In: *Journal of Artificial Intelligence Research* 16.1 (2002), pp. 321–357.

[40] N. V. Chawla, N. Japkowicz, and A. Kotcz. "Editorial: special issue on learning from imbalanced data sets". In: *ACM SIGKDD Explorations Newsletter* 6.1 (2004), pp. 1–6.

[41] G. W. Chodak, R. A. Thisted, G. S. Gerber, J.-E. Johansson, J. Adolfsson, G. W. Jones, G. D. Chisholm, B. Moskovitz, P. M. Livne, and J. Warner. "Results of conservative management of clinically localized prostate cancer". In: *New England Journal of Medicine* 330.4 (1994), pp. 242–248.

[42] F. K. Chun, A. Briganti, M. Graefen, C. Porter, F. Montorsi, A. Haese, V. Scattoni, L. Borden, T. Steuber, A. Salonia, et al. "Development and external validation of an extended repeat biopsy nomogram". In: *The Journal of Urology* 177.2 (2007), pp. 510–515.

[43] F. K. Chun, J. I. Epstein, V. Ficarra, S. J. Freedland, R. Montironi, F. Montorsi, S. F. Shariat, F. H. Schröder, and V. Scattoni. "Optimizing performance and interpretation of prostate biopsy: a critical analysis of the literature". In: *European Urology* 58.6 (2010), pp. 851–864.

[44] J. N. Cornu, G. Cancel Tassin, C. Egrot, C. Gaffory, F. Haab, and O. Cussenot. "Urine TMPRSS2: ERG fusion transcript integrated with PCA3 score, genotyping, and biological features are correlated to the results of prostatic biopsies in men at risk of prostate cancer". In: *The Prostate* 73.3 (2013), pp. 242–249.

[45] M. E. Cowen, L. K. Halasyamani, and M. W. Kattan. "Predicting life expectancy in men with clinically localized prostate cancer". In: *The Journal of Urology* 175.1 (2006), pp. 99–103.

[46] D. R. Cox. "Two further applications of a model for binary regression". In: *Biometrika* (1958), pp. 562–565.

[47] E. D. Crawford, K. O. Rove, E. J. Trabulsi, J. Qian, K. P. Drewnowska, J. C. Kaminetsky, T. K. Huisman, M. L. Bilowus, S. J. Freedman, W. L. Glover, et al. "Diagnostic performance of PCA3 to detect prostate cancer in men with increased prostate specific antigen: a prospective study of 1,962 cases". In: *The Journal of Urology* 188.5 (2012), pp. 1726–1731.

[48] J. A. Cruz and D. S. Wishart. "Applications of machine learning in cancer prediction and prognosis". In: *Cancer Informatics* 2 (2006), p. 59.

[49] J. E. Dennis Jr and J. J. Moré. "Quasi-Newton methods, motivation and theory". In: *SIAM review* 19.1 (1977), pp. 46–89.

[50] I. L. Deras, S. M. Aubin, A. Blase, J. R. Day, S. Koo, A. W. Partin, W. J. Ellis, L. S. Marks, Y. Fradet, H. Rittenhouse, et al. "PCA3: a molecular urine assay for predicting prostate biopsy outcome". In: *The Journal of Urology* 179.4 (2008), pp. 1587–1592.

[51] P. Domingos. "Metacost: A general method for making classifiers cost-sensitive". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 1999, pp. 155–164.

[52] B. Efron and R. Tibshirani. "Improvements on cross-validation: the 632+ bootstrap method". In: *Journal of the American Statistical Association* 92.438 (1997), pp. 548–560.

[53] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[54] D. U. Ekwueme, L. A. Stroud, and Y. Chen. "Peer Reviewed: Cost Analysis of Screening for, Diagnosing, and Staging Prostate Cancer Based on a Systematic Review of Published Studies". In: *Preventing Chronic Disease* 4.4 (2007).

[55] L. El Ghaoui and H. Lebret. "Robust solutions to least-squares problems with uncertain data". In: *SIAM Journal on matrix analysis and applications* 18.4 (1997), pp. 1035–1064.

[56] L. El Ghaoui, F. Oustry, and H. Lebret. "Robust solutions to uncertain semidefinite programs". In: *SIAM Journal on Optimization* 9.1 (1998), pp. 33–52.

[57] C. Elkan. "The foundations of cost-sensitive learning". In: *Proceedings of the International Joint Conference on Artificial Intelligence*. Vol. 17. 1. Citeseer. 2001, pp. 973–978.

[58] T. Evgeniou, M. Pontil, and T. Poggio. "Regularization networks and support vector machines". In: *Advances in Computational Mathematics* 13.1 (2000), pp. 1–50.

[59] X. Filella, L. Foj, M. Mila, J. M. Auge, R. Molina, and W. Jimenez. "PCA3 in the detection and management of early prostate cancer". In: *Tumor Biology* 34.3 (2013), pp. 1337–1347.

[60] Y. Fradet, F. Saad, A. Aprikian, J. Dessureault, M. Elhilali, C. Trudel, B. Masse, L. Piche, and C. Chypre. "uPM3, a new molecular urine test for the detection of prostate cancer". In: *Urology* 64.2 (2004), pp. 311–315.

[61] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Vol. 1. Springer, 2001.

[62] J. Friedman, T. Hastie, R. Tibshirani, et al. "Additive logistic regression: a statistical view of boosting". In: *The Annals of Statistics* 28.2 (2000), pp. 337–407.

[63] M. Fuchsjäger, A. Shukla-Dave, O. Akin, J. Barentsz, and H. Hricak. "Prostate cancer imaging". In: *Acta Radiologica* 49.1 (2008), pp. 107–120.

[64] M. P. van Gils, E. B. Cornel, D. Hessels, W. Peelen, J. A. Witjes, P. F. Mulders, H. G. Rittenhouse, and J. A. Schalken. "Molecular PCA3 diagnostics on prostatic fluid". In: *The Prostate* 67.8 (2007), pp. 881–887.

[65] M. C. Gittelman, B. Hertzman, J. Bailen, T. Williams, I. Koziol, R. J. Henderson, M. Efros, M. Bidair, and J. F. Ward. "PCA3 molecular urine test as a predictor of repeat prostate biopsy outcome in men with previous negative biopsies: a prospective multicenter clinical study". In: *The Journal of Urology* 190.1 (2013), pp. 64–69.

[66] W. J. Gradishar, B. O. Anderson, R. Balassanian, S. L. Blair, H. J. Burstein, A. Cyr, A. D. Elias, W. B. Farrar, A. Forero, S. H. Giordano, et al. "NCCN guidelines insights: breast cancer, version 1.2017". In: *Journal of the National Comprehensive Cancer Network* 15.4 (2017), pp. 433–451.

[67] R. Greiner, A. J. Grove, and D. Roth. "Learning cost-sensitive active classifiers". In: *Artificial Intelligence* 139.2 (2002), pp. 137–174.

[68] G. P. Haas, N. B. Delongchamps, R. F. Jones, V. Chandan, A. M. Serio, A. J. Vickers, M. Jumbelic, G. Threatte, R. Korets, H. Lilja, et al. "Needle biopsies on autopsy prostates: sensitivity of cancer detection based on true prevalence". In: *Journal of the National Cancer Institute* 99.19 (2007), pp. 1484–1489.

[69] A. Haese, A. de la Taille, H. Van Poppel, M. Marberger, A. Stenzl, P. F. Mulders, H. Huland, C.-C. Abbou, M. Remzi, M. Tinzl, et al. "Clinical utility of the PCA3 urine assay in European men scheduled for repeat biopsy". In: *European Urology* 54.5 (2008), pp. 1081–1088.

[70] H. Han, W.-Y. Wang, and B.-H. Mao. "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning". In: *Advances in Intelligent Computing* (2005), pp. 878–887.

[71] F. Harrell, K. L. Lee, and D. B. Mark. "Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors". In: *Statistics in Medicine* 15 (1996), pp. 361–387.

[72] P. Hart. "The condensed nearest neighbor rule". In: *IEEE Transactions on Information Theory* 14.3 (1968), pp. 515–516.

[73] H. He and E. A. Garcia. "Learning from imbalanced data". In: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (2009), pp. 1263–1284.

[74] A. Heidenreich, P. J. Bastian, J. Bellmunt, M. Bolla, S. Joniau, T. van der Kwast, M. Mason, V. Matveev, T. Wiegel, F. Zattoni, and N. Mottet. "EAU Guidelines on Prostate Cancer. Part II: Treatment of Advanced, Relapsing, and Castration-Resistant Prostate Cancer". In: *European Urology* 65.2 (2014), pp. 467–479. ISSN: 0302-2838. DOI: http://dx.doi.org/10.1016/j.eururo.2013.11.002. URL: http://www.sciencedirect.com/science/article/pii/S0302283813011998.

[75] A. Heidenreich, P. J. Bastian, J. Bellmunt, M. Bolla, S. Joniau, T. van der Kwast, M. Mason, V. Matveev, T. Wiegel, F. Zattoni, et al. "EAU guidelines on prostate cancer. Part 1: screening, diagnosis, and local treatment with curative intent update 2013". In: *European Urology* 65.1 (2014), pp. 124–137.

[76] D. Hessels, M. P. van Gils, O. van Hooij, S. A. Jannink, J. A. Witjes, G. W. Verhaegh, and J. A. Schalken. "Predictive value of PCA3 in urinary sediments in determining clinico-pathological characteristics of prostate cancer". In: *The Prostate* 70.1 (2010), pp. 10–16.

[77] D. Hessels, F. P. Smit, G. W. Verhaegh, J. A. Witjes, E. B. Cornel, and J. A. Schalken. "Detection of TMPRSS2-ERG fusion transcripts and prostate cancer antigen 3 in urinary sediments may improve diagnosis of prostate cancer". In: *Clinical Cancer Research* 13.17 (2007), pp. 5103–5108.

[78] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*. Vol. 398. John Wiley & Sons, 2013.

[79] H. Hricak, P. L. Choyke, S. C. Eberhardt, S. A. Leibel, and P. T. Scardino. "Imaging prostate cancer: a multidisciplinary perspective". In: *Radiology* 243.1 (2007), pp. 28–53.

[80] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. *A practical guide to support vector classification*. Tech. rep. National Taiwan University, Department of Computer Science, 2003. URL: `http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf` (visited on 06/04/2017).

[81] B. Huang, M. W.-M. Law, and P.-L. Khong. "Whole-body PET/CT scanning: estimation of radiation dose and cancer risk". In: *Radiology* 251.1 (2009), pp. 166–174.

[82] P. Hurley, J. Montie, A. Dhir, Y. Gao, B. Drabik, K. Lim, J. Curry, S. Linsell, A. Brachulis, K. Ghani, B. Denton, and D. Miller. "A statewide intervention to reduce the use of low value imaging among men with newly-diagnosed prostate cancer". In: *The Journal of Urology* 195.4 (2016), pp. 591–592.

[83] A. Jalilvand-Nejad, R. Shafaei, and H. Shahriari. "Robust optimization under correlated polyhedral uncertainty set". In: *Computers & Industrial Engineering* 92 (2016), pp. 82–94.

[84] A. Jemal, R. Siegel, J. Xu, and E. Ward. "Cancer statistics, 2010". In: *CA: A Cancer Journal for Clinicians* 60.5 (2010), pp. 277–300.

[85] J.-E. Johansson, O. Andren, S.-O. Andersson, P. W. Dickman, L. Holmberg, A. Magnuson, and H.-O. Adami. "Natural history of early, localized prostate cancer". In: *JAMA* 291.22 (2004), pp. 2713–2719.

[86] H. Kellerer, U. Pferschy, and D. Pisinger. *Knapsack Problems*. Berlin, Germany: Springer, 2004. ISBN: 978-3-540-24777-7.

[87] A. S. Kibel, J. P. Ciezki, E. A. Klein, C. A. Reddy, J. D. Lubahn, J. Haslag-Minoff, J. O. Deasy, J. M. Michalski, D. Kallogjeri, J. F. Piccirillo, et al. "Survival among men with clinically localized prostate cancer treated with radical prostatectomy or radiation therapy in the prostate specific antigen era". In: *The Journal of Urology* 187.4 (2012), pp. 1259–1265.

[88] G. Kimeldorf and G. Wahba. "Some results on Tchebycheffian spline functions". In: *Journal of Mathematical Analysis and Applications* 33.1 (1971), pp. 82–95.

[89] A. S. Kosinski and H. X. Barnhart. "Accounting for nonignorable verification bias in assessment of diagnostic tests". In: *Biometrics* 59.1 (2003), pp. 163–171.

[90] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis. "Machine learning applications in cancer prognosis and prediction". In: *Computational and Structural Biotechnology* 13 (2015), pp. 8–17.

[91] J. Laurikkala. "Improving identification of difficult small classes by balancing class distribution". In: *Proceedings of the Conference on Artificial Intelligence in Medicine in Europe*. Springer. 2001, pp. 63–66.

[92] B. Laxman, D. S. Morris, J. Yu, J. Siddiqui, J. Cao, R. Mehra, R. J. Lonigro, A. Tsodikov, J. T. Wei, S. A. Tomlins, et al. "A first-generation multiplex biomarker analysis of urine for the early detection of prostate cancer". In: *Cancer Research* 68.3 (2008), pp. 645–649.

[93] G. Lemaître, F. Nogueira, and C. K. Aridas. "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning". In: *Journal of Machine Learning Research* 18.17 (2017), pp. 1–5.

[94] G. H. Leyten, D. Hessels, S. A. Jannink, F. P. Smit, H. de Jong, E. B. Cornel, T. M. de Reijke, H. Vergunst, P. Kil, B. C. Knipscheer, et al. "Prospective multicentre evaluation of PCA3 and TMPRSS2-ERG gene fusions as diagnostic and prognostic urinary biomarkers for prostate cancer". In: *European Urology* 65.3 (2014), pp. 534–542.

[95] Y. Li, J. T. Y. Kwok, and Z. H. Zhou. "Cost-sensitive semi-supervised support vector machine". In: *Proceedings of the National Conference on Artificial Intelligence*. Vol. 1. 2010, p. 500.

[96] E. C. Lin. "Radiation risk from medical imaging". In: *Mayo Clinic Proceedings*. Vol. 85. 12. Elsevier. 2010, pp. 1142–1146.

[97] R. J. Little. "A test of missing completely at random for multivariate data with missing values". In: *Journal of the American Statistical Association* 83.404 (1988), pp. 1198–1202.

[98] A. Liu, G. Jun, and J. Ghosh. "Spatially cost-sensitive active learning". In: *Proceedings of the SIAM International Conference on Data Mining*. SIAM. 2009, pp. 814–825.

[99] L. Liu, A. L. Coker, X. L. Du, J. N. Cormier, C. E. Ford, and S. Fang. "Long-term survival after radical prostatectomy compared to other treatments in older men with local/regional prostate cancer". In: *Journal of Surgical Oncology* 97.7 (2008), pp. 583–591.

[100] X. Y. Liu and Z. H. Zhou. "The influence of class imbalance on cost-sensitive learning: An empirical study". In: *Proceedings of the International Conference on Data Mining*. IEEE. 2006, pp. 970–974.

[101] Y. Lotan, A. Q. Haddad, D. N. Costa, I. Pedrosa, N. M. Rofsky, and C. G. Roehrborn. "Decision analysis model comparing cost of multiparametric magnetic resonance imaging vs. repeat biopsy for detection of prostate cancer in men with prior negative findings on biopsy". In: *Urologic Oncology: Seminars and Original Investigations*. Vol. 33. 6. Elsevier. 2015, 266–e9.

[102] M. Maalouf, T. B. Trafalis, and I. Adrianto. "Kernel logistic regression using truncated Newton method". In: *Computational Management Science* 8.4 (2011), pp. 415–428.

[103] M. A. Maloof. "Learning when data sets are imbalanced and when costs are unequal and unknown". In: *Proceedings of the ICML Workshop on Learning from Imbalanced Datasets II*. Vol. 2. 2003, pp. 2–1.

[104] D. D. Margineantu. "Active cost-sensitive learning". In: *Proceedings of the International Joint Conference on Artificial Intelligence*. Vol. 5. 2005, pp. 1622–1623.

[105] L. S. Marks, Y. Fradet, I. L. Deras, A. Blase, J. Mathis, S. M. Aubin, A. T. Cancio, M. Desaulniers, W. J. Ellis, H. Rittenhouse, et al. "PCA3 molecular urine assay for prostate cancer in men undergoing repeat biopsy". In: *Urology* 69.3 (2007), pp. 532–535.

[106] H. Masnadi-Shirazi and N. Vasconcelos. "Risk minimization, probability elicitation, and cost-sensitive SVMs." In: *Proceedings of the International Conference on Machine learning*. 2010, pp. 759–766.

[107] K. McCarthy, B. Zabar, and G. Weiss. "Does cost-sensitive learning beat sampling for classifying rare classes?" In: *Proceedings of the International Workshop on Utility-based Data Mining*. ACM. 2005, pp. 69–77.

[108] S. Merdan, C. Barnett, T. B. Denton, J. E. Montie, and D. Miller. "Data Analytics for Optimal Detection of Metastatic Prostate Cancer". In: *Submitted* (2017).

[109] S. Merdan, S. A. Tomlins, C. L. Barnett, T. M. Morgan, J. E. Montie, J. T. Wei, and B. T. Denton. "Assessment of long-term outcomes associated with urinary prostate cancer antigen 3 and TMPRSS2: ERG gene fusion at repeat biopsy". In: *Cancer* 121.22 (2015), pp. 4071–4079.

[110] S. Merdan, P. R. Womble, D. C. Miller, C. Barnett, Z. Ye, S. M. Linsell, J. E. Montie, and B. T. Denton. "Toward better use of bone scans among men with early-stage prostate cancer". In: *Urology* 84.4 (2014), pp. 793–798.

[111] M. E. Miller, C. D. Langefeld, W. M. Tierney, S. L. Hui, and C. J. McDonald. "Validation of probabilistic predictions". In: *Medical Decision Making* 13.1 (1993), pp. 49–57.

[112] K. G. M. Moons, A. P. Kengne, D. E. Grobbee, P. Royston, Y. Vergouwe, D. G. Altman, and M. Woodward. "Risk prediction models: II. External validation, model updating, and impact assessment". In: *Heart* 98.9 (2012), pp. 691–698. ISSN: 1355-6037. DOI: 10.1136/heartjnl-2011-301247. eprint: http://heart.bmj.com/content/98/9/691.full.pdf. URL: http://heart.bmj.com/content/98/9/691.

[113] K. G. M. Moons, A. P. Kengne, M. Woodward, P. Royston, Y. Vergouwe, D. G. Altman, and D. E. Grobbee. "Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker". In: *Heart* 98.9 (2012), pp. 683–690. ISSN: 1355-6037. DOI: 10.1136/heartjnl-2011-301246. eprint: http://heart.bmj.com/content/98/9/683.full.pdf. URL: http://heart.bmj.com/content/98/9/683.

[114] K. G. Moons, D. G. Altman, J. B. Reitsma, J. P. Ioannidis, P. Macaskill, E. W. Steyerberg, A. J. Vickers, D. F. Ransohoff, and G. S. Collins. "Transparent Reporting of a multivariable prediction model for individual Prognosis or diagnosis (TRIPOD): Explanation and ElaborationThe TRIPOD Statement: explanation and elaboration". In: *Annals of Internal Medicine* 162.1 (2015), W1–W73.

[115] N. Mottet, J. Bellmunt, E. Briers, E. U. Association, et al. "Guidelines on prostate cancer". In: *European Urology* 65.1 (2014), pp. 124–37.

[116] H. Nakanishi, J. Groskopf, H. A. Fritsche, V. Bhadkamkar, A. Blase, S. V. Kumar, J. W. Davis, P. Troncoso, H. Rittenhouse, and R. J. Babaian. "PCA3 molecular urine assay correlates with prostate cancer tumor volume: implication in selecting candidates for active surveillance". In: *The Journal of Urology* 179.5 (2008), pp. 1804–1810.

[117] NCCN. *National Comprehensive Cancer Network Clinical Guidelines*. `https://www.nccn.org/professionals/physician_gls/f_guidelines.asp`. 2014.

[118] A. R. Padhani, F. E. Lecouvet, N. Tunariu, D.-M. Koh, F. De Keyzer, D. J. Collins, E. Sala, S. Fanti, H. A. Vargas, G. Petralia, et al. "Rationale for modernising imaging in advanced prostate cancer". In: *European Urology Focus* (2016).

[119] R. P. Pal, N. U. Maitra, J. K. Mellon, and M. A. Khan. "Defining prostate cancer risk before prostate biopsy". In: *Urologic Oncology: Seminars and Original Investigations*. Vol. 31. 8. Elsevier. 2013, pp. 1408–1418.

[120] PCPT. *Prostate Cancer Prevention Trial Risk Calculator Version 2.0*. `http://myprostatecancerrisk.com/`. 2006.

[121] M. S. Pepe. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, 2003.

[122] S. Perdona, V. Cavadas, G. Di Lorenzo, R. Damiano, G. Chiappetta, P. Del Prete, R. Franco, G. Azzarito, S. Scala, C. Arra, et al. "Prostate cancer detection in the grey area of prostatespecific antigen below 10 ng/ml: headtohead comparison of the updated PCPT calculator and Chun's nomogram, two risk estimators incorporating prostate cancer antigen 3". In: *European Urology* 59.1 (2011), pp. 81–87.

[123] A. Pettersson, R. E. Graff, S. R. Bauer, M. J. Pitt, R. T. Lis, E. C. Stack, N. E. Martin, L. Kunz, K. L. Penney, A. H. Ligon, et al. "The TMPRSS2: ERG rearrangement, ERG expression, and prostate cancer outcomes: a cohort study and meta-analysis". In: *Cancer Epidemiology and Prevention Biomarkers* 21.9 (2012), pp. 1497–1509.

[124] D. A. Pierce and D. L. Preston. "Radiation-related cancer risks at low doses among atomic bomb survivors". In: *Radiation research* 154.2 (2000), pp. 178–186.

[125] G. Ploussard, X. Durand, E. Xylinas, S. Moutereau, C. Radulescu, A. Forgue, N. Nicolaiew, S. Terry, Y. Allory, S. Loric, et al. "Prostate cancer antigen 3 score accurately predicts tumour volume and might help in selecting prostate cancer patients for active surveillance". In: *European Urology* 59.3 (2011), pp. 422–429.

[126] G. Ploussard, A. Haese, H. Van Poppel, M. Marberger, A. Stenzl, P. F. Mulders, H. Huland, L. Bastien, C.-C. Abbou, M. Remzi, et al. "The prostate cancer gene 3 (PCA3) urine test in men with previous negative biopsies: does free-to-total prostate-specific antigen ratio influence the performance of the PCA3 score in predicting positive biopsies?" In: *BJU International* 106.8 (2010), pp. 1143–1147.

[127] R. Pockett, D. Castellano, P. McEwan, A. Oglesby, B. Barber, and K. Chung. "The hospital burden of disease associated with bone metastases and skeletal-related events in patients with breast cancer, lung cancer, or prostate cancer in Spain". In: *European Journal of Cancer Care* 19.6 (2010), pp. 755–760.

[128] K. N. Prasad, W. C. Cole, and G. M. Hasse. "Health risks of low dose ionizing radiation in humans: a review". In: *Experimental Biology and Medicine* 229.5 (2004), pp. 378–382.

[129] A. Prékopa. *Stochastic programming*. Vol. 324. Springer Science & Business Media, 2013.

[130] Z. Qi, Y. Tian, Y. Shi, and X. Yu. "Cost-Sensitive Support Vector Machine for Semi-Supervised Learning". In: *Procedia Computer Science* 18 (2013). International Conference on Computational Science, pp. 1684–1689.

[131] Z. Qin, S. Zhang, L. Liu, and T. Wang. "Cost-sensitive semi-supervised classification using CS-EM". In: *Proceedings of the IEEE International Conference on Computer and Information Technology*. IEEE. 2008, pp. 131–136.

[132] J. Raja, N. Ramachandran, G. Munneke, and U. Patel. "Current status of transrectal ultrasoundguided prostate biopsy in the diagnosis of prostate cancer". In: *Clinical Radiology* 61.2 (2006), pp. 142–153.

[133] R. Risko, S. Merdan, P. R. Womble, C. Barnett, Z. Ye, S. M. Linsell, J. E. Montie, D. C. Miller, and B. T. Denton. "Clinical predictors and recommendations for staging computed tomography scan among men with prostate cancer". In: *Urology* 84.6 (2014), pp. 1329–1334.

[134] A. W. Roddam, M. J. Duffy, F. C. Hamdy, A. M. Ward, J. Patnick, C. P. Price, J. Rimmer, C. Sturgeon, P. White, and N. E. Allen. "Use of prostate-specific antigen (PSA) isoforms for the detection of prostate cancer in men with a PSA level of 2–10 ng/ml: systematic review and meta-analysis". In: *European Urology* 48.3 (2005), pp. 386–399.

[135] M. J. Roobol, F. H. Schroder, P. van Leeuwen, T. Wolters, R. C. van den Bergh, G. J. van Leenders, and D. Hessels. "Performance of the prostate cancer antigen 3 (PCA3) gene and prostatespecific antigen in prescreened men: exploring the value of PCA3 for a firstline diagnostic test". In: *European Urology* 58.4 (2010), pp. 475–481.

[136] J. Rubio Briones, A. Fernandez Serra, M. Ramirez, L. Rubio, A. Collado, J. Casanova, A. Gomez Ferrer, J. Ricos, J. Monros, R. Dumont, et al. "Outcomes of expanded use of PCA3 testing in a Spanish population with clinical suspicion of prostate cancer". In: *Actas Urologicas Espanolas* 35.10 (2011), pp. 589–596.

[137] J. Ruiz-Aragon and S. Marquez-Pelaez. "Assessment of the PCA3 test for prostate cancer diagnosis: a systematic review and meta-analysis". In: *Actas Urológicas Españolas* 34.4 (2010), pp. 346–355.

[138] A. P. Ruszczynski and A. Shapiro. *Stochastic programming (Handbooks in Operations Research and Management Science)*. Vol. 10. Elsevier Amsterdam, 2003.

[139] M. Salagierski and J. A. Schalken. "Molecular diagnosis of prostate cancer: PCA3 and TMPRSS2: ERG gene fusion". In: *The Journal of Urology* 187.3 (2012), pp. 795–801.

[140] S. S. Salami, F. Schmidt, B. Laxman, M. M. Regan, D. S. Rickman, D. Scherr, G. Bueti, J. Siddiqui, S. A. Tomlins, J. T. Wei, et al. "Combining urinary detection of TMPRSS2: ERG and PCA3 with serum PSA to predict diagnosis of prostate cancer". In: *Urologic Oncology*. Vol. 31. 5. Elsevier. 2013, pp. 566–571.

[141] R. E. Schapire. "The boosting approach to machine learning: An overview". In: *Nonlinear Estimation and Classification*. Springer, 2003, pp. 149–171.

[142] L. E. Schnipper, T. J. Smith, D. Raghavan, D. W. Blayney, P. A. Ganz, T. M. Mulvey, and D. S. Wollins. "American Society of Clinical Oncology identifies five key opportunities to improve care and reduce costs: the top five list for oncology". In: *J Clin Oncol* 30.14 (2012), pp. 1715–1724.

[143] F. H. Schröder, J. Hugosson, M. J. Roobol, T. L. Tammela, S. Ciatto, V. Nelen, M. Kwiatkowski, M. Lujan, H. Lilja, M. Zappa, et al. "Screening and prostate-cancer mortality in a randomized European study". In: *New England Journal of Medicine* 360.13 (2009), pp. 1320–1328.

[144] C. Scott. "Calibrated asymmetric surrogate losses". In: *Electronic Journal of Statistics* 6 (2012), pp. 958–992.

[145] C. Seitz, S. Palermo, and B. Djavan. "Prostate biopsy". In: *The Italian Journal of Urology and Nephrology* 55.4 (2003), pp. 205–218.

[146] S. B. Shappell, J. Fulmer, D. Arguello, B. S. Wright, J. R. Oppenheimer, and M. J. Putzi. "PCA3 urine mRNA testing for prostate carcinoma: patterns of use by community urologists and assay performance in reference laboratory setting". In: *Urology* 73.2 (2009), pp. 363–368.

[147] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge University Press, 2004.

[148] V. Sindhwani, P. Niyogi, M. Belkin, and S. Keerthi. "Linear manifold regularization for large scale semi-supervised learning". In: *Proceedings of the ICML Workshop on Learning with Partially Classified Training Data*. Vol. 28. 2005.

[149] R. Smith-Bindman, J. Lipson, R. Marcus, K.-P. Kim, M. Mahesh, R. Gould, A. B. De González, and D. L. Miglioretti. "Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer". In: *Archives of Internal Medicine* 169.22 (2009), pp. 2078–2086.

[150] A. L. Soyster. "Convex Programming with Set-Inclusive Constraints and Applications to Inexact Linear Programming". In: *Operations Research* 21.5 (1973), pp. 1154–1157. ISSN: 0030364X, 15265463. URL: `http://www.jstor.org/stable/168933`.

[151] C. Stephan, K. Jung, A. Semjonow, K. Schulze Forster, H. Cammann, X. Hu, H. A. Meyer, M. Bogemann, K. Miller, and F. Friedersdorff. "Comparative assessment of urinary prostate cancer antigen 3 and TMPRSS2: ERG gene fusion with the serum [- 2] proprostatespecific antigenbased prostate health index for detection of prostate cancer". In: *Clinical Chemistry* 59.1 (2013), pp. 280–288.

[152] A. J. Stephenson, M. W. Kattan, J. A. Eastham, F. J. Bianco Jr, O. Yossepowitch, A. J. Vickers, E. A. Klein, D. P. Wood, and P. T. Scardino. "Prostate cancerspecific mortality after radical prostatectomy for patients treated in the prostate-specific antigen era". In: *Journal of Clinical Oncology* 27.26 (2009), pp. 4300–4305.

[153] E. W. Steyerberg. *Clinical prediction models: a practical approach to development, validation, and updating*. Springer Science & Business Media, 2008.

[154] E. W. Steyerberg, F. E. Harrell, G. J. Borsboom, M. Eijkemans, Y. Vergouwe, and J. D. F. Habbema. "Internal validation of predictive models: efficiency of some procedures for logistic regression analysis". In: *Journal of Clinical Epidemiology* 54.8 (2001), pp. 774–781.

[155] Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang. "Cost-sensitive boosting for classification of imbalanced data". In: *Pattern Recognition* 40.12 (2007), pp. 3358–3378.

[156] S. D. Sweat, E. J. Bergstralh, J. Slezak, M. L. Blute, and H. Zincke. "Competing risk analysis after radical prostatectomy for clinically nonmetastatic prostate adenocarcinoma according to clinical Gleason score and patient age". In: *The Journal of Urology* 168.2 (2002), pp. 525–529.

[157] A. de la Taille, J. Irani, M. Graefen, F. Chun, T. de Reijke, P. Kil, P. Gontero, A. Mottaz, and A. Haese. "Clinical evaluation of the PCA3 assay in guiding initial biopsy decisions". In: *The Journal of Urology* 185.6 (2011), pp. 2119–2125.

[158] M. K. Terris. "Sensitivity and specificity of sextant biopsies in the detection of prostate cancer: preliminary report". In: *Urology* 54.3 (1999), pp. 486–489.

[159] A. Tewari, C. C. Johnson, G. Divine, E. D. Crawford, E. J. Gamito, R. Demers, and M. Menon. "Long-term survival probability in men with clinically localized prostate cancer: a case-control, propensity modeling study stratified by race, age, treatment and comorbidities". In: *The Journal of Urology* 171.4 (2004), pp. 1513–1519.

[160] I. M. Thompson, D. P. Ankerst, C. Chi, P. J. Goodman, C. M. Tangen, M. S. Lucia, Z. Feng, H. L. Parnes, and C. A. Coltman Jr. "Assessing prostate cancer risk: results from the Prostate Cancer Prevention Trial". In: *Journal of the National Cancer Institute* 98.8 (2006), pp. 529–534.

[161] I. Thompson, J. Thrasher, G. Aus, A. Burnett, E. Canby-Hagino, M. Cookson, A. D'Amico, R. Dmochowski, D. Eton, J. Forman, S. Goldenberg, J. Hernandez, C. Higano, S. Kraus, J. Moul, and C. Tangen. "Guideline for the Management of Clinically Localized Prostate Cancer: 2007 Update". In: *Journal of Urology* 177.6 (2007), pp. 2106–2131. ISSN: 0022-5347. DOI: 10.1016/j.juro.2007.03.003.

[162] K. M. Ting. "An instance-weighting method to induce cost-sensitive trees". In: *IEEE Transactions on Knowledge and Data Engineering* 14.3 (2002), pp. 659–665.

[163] F. Tokan, N. Türker, and T. Yıldırım. "ROC analysis as a useful tool for performance evaluation of artificial neural networks". In: *Artificial Neural Networks–ICANN 2006* (2006), pp. 923–931.

[164] B. Tombal, G. L. Andriole, A. de la Taille, P. Gontero, A. Haese, M. Remzi, M. Speakman, L. Smets, and H. Stoevelaar. "Clinical judgment versus biomarker prostate cancer gene 3: which is best when determining the need for repeat prostate biopsy?" In: *Urology* 81.5 (2013), pp. 998–1004.

[165] B. Tombal and F. Lecouvet. "Modern detection of prostate cancer's bone metastasis: is the bone scan era over?" In: *Advances in Urology* 2012 (2012).

[166] I. Tomek. "Two Modifications of CNN." In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-6.11 (1976), pp. 769–772.

[167] S. A. Tomlins, S. M. Aubin, J. Siddiqui, R. J. Lonigro, L. Sefton Miller, S. Miick, S. Williamsen, P. Hodge, J. Meinke, A. Blase, et al. "Urine TMPRSS2: ERG fusion transcript stratifies prostate cancer risk in men with elevated serum PSA". In: *Science Translational Medicine* 3.94 (2011), 94ra72–94ra72.

[168] S. A. Tomlins, J. R. Day, R. J. Lonigro, D. H. Hovelson, J. Siddiqui, L. P. Kunju, R. L. Dunn, S. Meyer, P. Hodge, J. Groskopf, et al. "Urine TMPRSS2: ERG plus PCA3 for individualized prostate cancer risk assessment". In: *European Urology* 70.1 (2016), pp. 45–53.

[169] V. Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2013.

[170] K. Veropoulos, C. Campbell, and N. Cristianini. "Controlling the sensitivity of support vector machines". In: *Proceedings of the International Joint Conference on Artificial Intelligence*. 1999, pp. 55–60.

[171] U. Von Luxburg. "A tutorial on spectral clustering". In: *Statistics and Computing* 17.4 (2007), pp. 395–416.

[172] J. Wade, D. J. Rosario, R. C. Macefield, K. N. Avery, C. E. Salter, M. L. Goodwin, J. M. Blazeby, J. A. Lane, C. Metcalfe, D. E. Neal, et al. "Psychological impact of prostate biopsy: physical symptoms, anxiety, and depression". In: *Journal of Clinical Oncology* 31.33 (2013), pp. 4235–4241.

[173] J. Walz, A. Gallina, F. Saad, F. Montorsi, P. Perrotte, S. F. Shariat, C. Jeldres, M. Graefen, F. Benard, M. McCormack, et al. "A nomogram predicting 10year life expectancy in candidates for radical prostatectomy or radiotherapy for prostate cancer". In: *Journal of Clinical Oncology* 25.24 (2007), pp. 3576–3581.

[174] R. Wang, A. M. Chinnaiyan, R. L. Dunn, K. J. Wojno, and J. T. Wei. "Rational approach to implementation of prostate cancer antigen 3 into clinical care". In: *Cancer* 115.17 (2009), pp. 3879–3886.

[175] G. M. Weiss. "Mining with rarity: a unifying framework". In: *ACM SIGKDD Explorations Newsletter* 6.1 (2004), pp. 7–19.

[176] D. L. Wilson. "Asymptotic properties of nearest neighbor rules using edited data". In: *IEEE Transactions on Systems, Man, and Cybernetics* 2.3 (1972), pp. 408–421.

[177] S. R. Xie, D. R. Bish, E. K. Bish, A. D. Slonim, and S. L. Stramer. "Safety and waste considerations in donated blood screening". In: *European Journal of Operational Research* 217.3 (2012), pp. 619–632.

[178] C. Yong, E. Onukwugha, and C. D. Mullins. "Clinical and economic burden of bone metastasis and skeletal-related events in prostate cancer". In: *Current Opinion in Oncology* 26.3 (2014), pp. 274–283.

[179] A. Young, N. Palanisamy, J. Siddiqui, D. P. Wood, J. T. Wei, A. M. Chinnaiyan, L. P. Kunju, and S. A. Tomlins. "Correlation of urine TMPRSS2: ERG and PCA3 to ERG+ and total prostate cancer burden". In: *American Journal of Clinical Pathology* 138.5 (2012), pp. 685–696.

[180] X.-H. Zhou. "Correcting for verification bias in studies of a diagnostic test's accuracy". In: *Statistical Methods in Medical Research* 7.4 (1998), pp. 337–353.

[181] X.-H. Zhou, D. K. McClish, and N. A. Obuchowski. *Statistical methods in diagnostic medicine*. Vol. 569. John Wiley & Sons, 2009.

[182] Z. H. Zhou and M. Li. "Semi-supervised learning by disagreement". In: *Knowledge and Information Systems* 24.3 (2010), pp. 415–439.

[183] Z. H. Zhou and X. Y. Liu. "Training cost-sensitive neural networks with methods addressing the class imbalance problem". In: *IEEE Transactions on Knowledge and Data Engineering* 18.1 (2006), pp. 63–77.

[184] J. Zhu and T. Hastie. "Kernel logistic regression and the import vector machine". In: *Journal of Computational and Graphical Statistics* 14.1 (2005).

[185] X. Zhu. *Semi-supervised learning literature survey*. Tech. rep. Computer Science, University of Wisconsin-Madison, 2007. URL: `http://pages.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf` (visited on 06/05/2017).

[186] X. Zhu and A. B. Goldberg. "Introduction to semi-supervised learning". In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 3.1 (2009), pp. 1–130.