Data Analytics for Optimal Detection of Metastatic Prostate Cancer

Selin Merdan, Christine Barnett, Brian T. Denton
Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI, 48109
smerdan@umich.edu, clbarnet@umich.edu, btdenton@umich.edu

James E. Montie, David C. Miller
Department of Urology, University of Michigan, Ann Arbor, MI, 48109
Michigan Urological Surgery Improvement Collaborative, Ann Arbor, MI, 48109
jmontie@umich.edu, dcmiller@umich.edu

We used data-analytics approaches to develop, calibrate, and validate predictive models, to help urologists in a large state-wide collaborative make prostate cancer staging decisions on the basis of individual patient risk factors. The models were validated using statistical methods based on bootstrapping and evaluation on out-of-sample data. These models were used to design guidelines that optimally weigh the benefits and harms of radiological imaging for detection of metastatic prostate cancer. The Michigan Urological Surgery Improvement Collaborative, a state-wide medical collaborative, implemented these guidelines, which were predicted to reduce unnecessary imaging by more than 40% and limit the percentage of patients with missed metastatic disease to be less than 1%. The effects of the guidelines were measured post-implementation to confirm their impact on reducing unnecessary imaging across the state of Michigan.

Key words: healthcare; prostate cancer: radiographic staging; semi-supervised learning; class imbalance problem; cost-sensitive learning; verification bias

1. Introduction

Prostate cancer is the most common cancer among men. It has been estimated that in 2017 there will be more than 160,000 new cases of prostate cancer 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 prostate cancer 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.

There are harms associated with both over- and under-imaging. Under-imaging results in patients' metastatic prostate cancer 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; 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 prostate cancer (Mottet et al. (2014), Thompson et al. (2007), NCCN (2014), Briganti et al. (2010), Heidenreich et al. (2014), Carroll et al. (2013)). Thus, there exists persistent variation in utilization among urologists, including unnecessary imaging in patients at low risk for metastatic disease and potentially incomplete staging of patients at high risk. 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 (see `http://musicurology.com/`).

We used a collection of methods including statistics, machine learning, and optimization methods that we collectively refer to as *data-analytics* methods. The key contributions of this article are as follows:

- *Risk Prediction Models for Metastatic Prostate Cancer.* We develop 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 were 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 optimization and machine learning methods to design classification rules that distinguish metastatic patients from cancer-free patients. To our knowledge, this is the first study to employ classification modeling techniques in the detection of metastatic prostate 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.

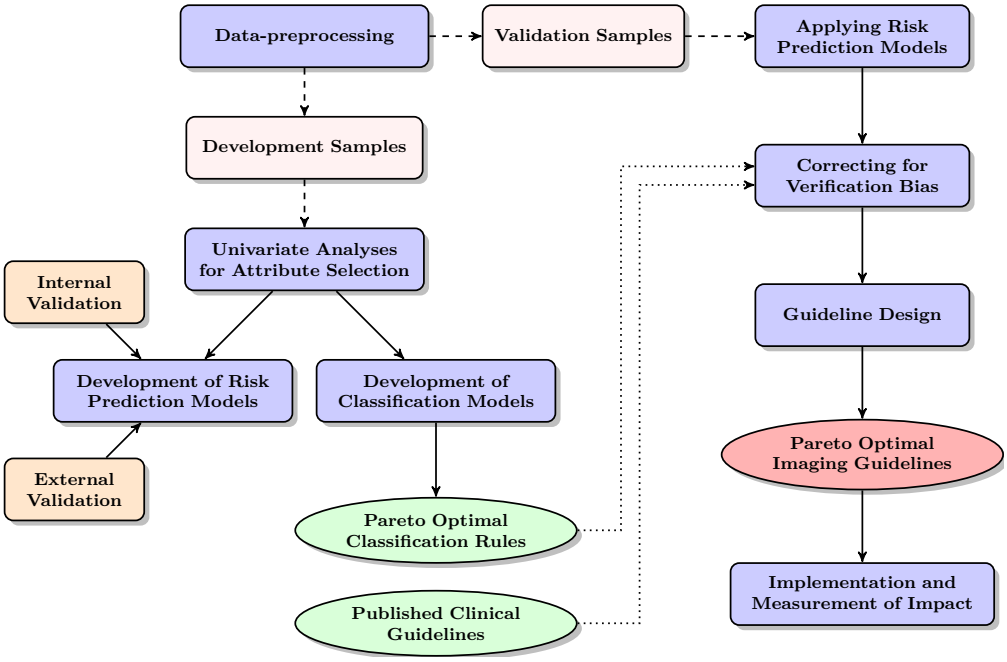
- *Bias-corrected Performance of Imaging Guidelines.* Because not all men with newly-diagnosed prostate cancer underwent imaging, we applied statistical methods to mitigate bias to evaluate the diagnostic accuracy of imaging guidelines for detection of metastatic disease. Our definition of imaging guidelines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

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

Figure 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 study is structured as follows. Section 2 describes the methodological approach for development and validation of risk prediction models, and proper measures for evaluating prediction performance. Section 3 reviews the challenges of classification modeling in imbalanced observational health data and describes our proposed algorithm for cost-sensitive semi-supervised learning. Section 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 5 describes the implementation process and the impact of our work based on post-implementation analysis. Section 6 highlights our main conclusions and states some points for future research.

Figure 1 Research framework illustrating the major steps from data preprocessing to implementation and measurement of impact.



2. Risk Prediction Models for Metastatic Prostate Cancer

In order 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.

2.1. 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.

Prostate cancer 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. Gleason score 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 prostate cancer. Clinical T stage is part of the TNM staging system for prostate cancer 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 prostate cancer seen at 19 MUSIC practices in Michigan from March 2012 through June 2013, and among this group, $416\,(27.39\%)$ 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 prostate cancer 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.

We performed univariate and multivariate analyses to examine the association between imaging outcomes and all routinely available clinical variables in imaged patients. 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. We used a logarithmic transformation of PSA scores since the distribution of PSA was highly skewed.

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, Gleason score, 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 logistic regression (LR) models to estimate the probability of a positive imaging outcome. The 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 learned via *maximum likelihood estimation* (MLE) by minimizing the conditional negative log-likelihood:

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

to obtain well-calibrated predicted probabilities.

2.3. Statistical Validation

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 (Harrell et al. (1996), Efron and Tibshirani (1997)).

Since internal validation has limitations in determining the generalizability of a predictive model (Bleeker et al. (2003)), 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.

2.3.1. 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 characteristic (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 (Tokan et al. (2006)). The larger the AUC the better the performance of the classification model.

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 estimating the slope of the linear predictor of the LR model, known as the *calibration slope* (Miller et al. (1993)). 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, $\text{logit}(\mathbb{P}(y = 1)) = \alpha + \beta_{calibration}\text{LP}$)(Cox (1958)). 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 (Harrell et al. (1996), Miller et al. (1993)).

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.

2.3.2. 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 Efron and Tibshirani (1994). We used this approach to internally validate the model calibration and discrimination.

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, $P_{validated}$.

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

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

- Draw a random bootstrap sample of n patients from the development sample with replacement.
- Fit the logistic regression model to the bootstrap sample and measure the apparent performance in the same sample, $P_{bootstrap}(i)$.
- Apply the bootstrap model to the development sample and estimate the test performance of this bootstrap model, $P_{test}(i)$.
- Calculate an estimate of the optimism, $o(i) = P_{bootstrap}(i) - P_{test}(i)$.

Estimate the expected optimism:

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

return $P_{validated} = P_{apparent} - Optimism$.

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 2.3.1.

2.4. Statistical Validation Results

Based on the approach described in Section 2.3.2, we calculated the expected optimism for the AUC, Brier score, and calibration slope (Table 1). 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.

Table 1 Bootstrap results for the development samples.

	Development samples	
	BS (n = 416) mean ± $\mathbf{SE_{bootstrap}}$	**CT scan (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 ($SE_{bootstrap}$).

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 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 positive outcome with evidence for bone metastasis, and for CT 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 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.

Table 2 Internal and external validation results of the risk prediction models.

	Development samples		**Validation samples**	
	BS (n = 416)	**CT scan (n = 643)**	**BS (n = 664)**	**CT scan (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.

The calibration plots in Figure 2 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.

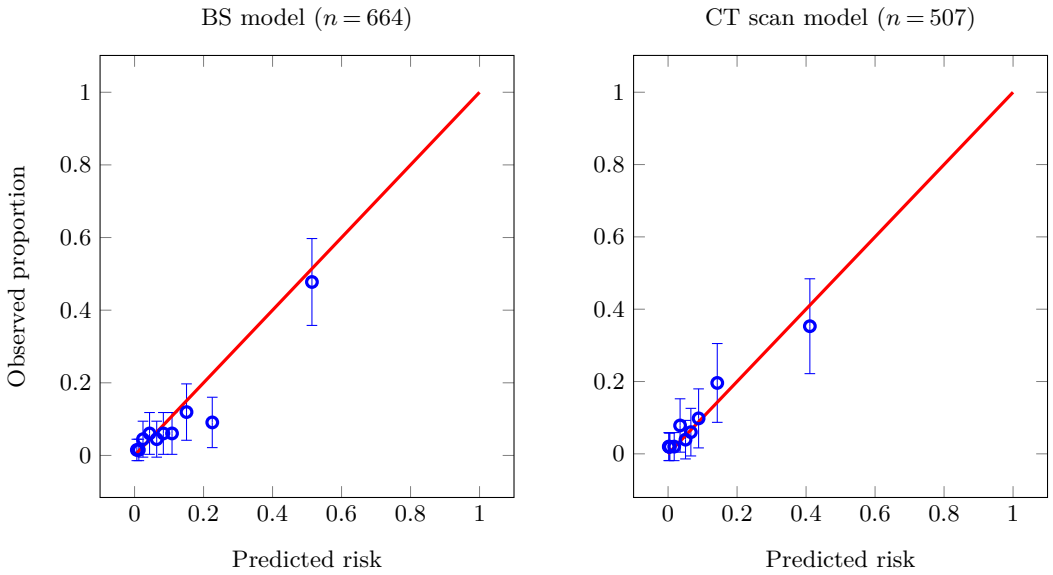
3. Classification Modeling for Metastatic Cancer Detection

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.

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

Figure 2 Calibration plots for BS and CT scan risk prediction models based on the validation samples.



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 article: *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 (Zhu (2007), Chapelle et al. (2010), Zhu and Goldberg (2009), Zhou and Li (2010)). 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 (Chawla et al. (2004), Weiss (2004), He and Garcia (2009)). Cost-sensitive learning aims to make the optimal decision that minimizes the total misclassification cost (Maloof (2003), Ting (2002), Domingos (1999), Elkan (2001), Masnadi-Shirazi and Vasconcelos (2010)). Several studies have shown that cost-sensitive methods demonstrated better performance than sampling methods in certain application domains (McCarthy et al. (2005), Liu and Zhou (2006), Zhou and Liu (2006), Sun et al. (2007)).

The use of unlabeled data in cost-sensitive learning has attracted growing attention and many techniques have been developed (Greiner et al. (2002), Margineantu (2005), Qin et al. (2008), Liu

et al. (2009), Li et al. (2010), Qi et al. (2013)). 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 Kourou et al. (2015) and Cruz and Wishart (2006)). In this article, 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 (Belkin et al. (2006)).

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 (Belkin et al. (2006), Evgeniou et al. (2000)). 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:

$$\min_{f \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} L(y_i, f(\mathbf{x}_i)) + \gamma_{\mathcal{H}} \|f\|_{\mathcal{H}}^2 \tag{3.2.1}$$

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 (Shawe-Taylor and Cristianini (2004)). 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 (Kimeldorf and Wahba (1971)), the minimizer $f^*(\mathbf{x})$ of (3.2.1) has the form:

$$f^*(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i^* \mathbf{K}(\mathbf{x}, \mathbf{x}_i) \tag{3.2.2}$$

As a consequence, (3.2.1) 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 machine (SVM) and regularized least squares (Belkin et al. (2006)).

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, $\langle \boldsymbol{\Phi}(\mathbf{x}) , \boldsymbol{\Phi}(\mathbf{x}') \rangle = \mathbf{K}(\mathbf{x}, \mathbf{x}')$.

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 (3.2.1), 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: $f(\mathbf{x}) = \Phi(\mathbf{x})\boldsymbol{\beta}$ (Zhu and Hastie (2005), Maalouf et al. (2011)). The optimization problem becomes as follows:

$$\min_{\boldsymbol{\beta} \in \mathcal{H}} \sum_{i=1}^{l} \log(1 + \exp(-y_i \langle \boldsymbol{\beta}, \Phi(\mathbf{x}_j) \rangle)) + \frac{\lambda}{2} \|\boldsymbol{\beta}\|^2, \tag{3.2.3}$$

where $\boldsymbol{\beta} \in \mathcal{H}$ is the parameter we want to estimate. By (3.2.2) and the kernel trick, the minimizer of (3.2.3) admits a representation of the form $\boldsymbol{\beta} = \sum_{i=1}^{l} \alpha_i \Phi(\mathbf{x}_i)$. Thus, we can write (3.2.3) as:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^l} \sum_{i=1}^{l} \log(1 + \exp(-y_i (\mathbf{K}\boldsymbol{\alpha})_i)) + \frac{\lambda}{2} \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} \tag{3.2.4}$$

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 Belkin et al. (2006), which extends the classical framework of regularization given in (3.2.1) 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 (Sindhwani et al. (2005)).

The concept underlying this new regularization comes from *spectral clustering*, which is one of the most popular clustering algorithms (Von Luxburg (2007)). 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:

$$f^* = \arg\min_{f\in\mathcal{H}} \frac{1}{l}\sum_{i=1}^{l} L\left(y_i, f(\mathbf{x}_i)\right) + \gamma_{\mathcal{H}}\|f\|_{\mathcal{H}}^2 + \gamma_{\mathcal{M}}\mathbf{f}^{\mathbf{T}}\mathbf{L}\mathbf{f} \tag{3.2.5}$$

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 decision 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 article, 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:

$$L_\delta = \delta \mathbb{1}_{\{y=1\}} L_1(f(\mathbf{x})) + (1-\delta)\mathbb{1}_{\{y=-1\}} L_{-1}(f(\mathbf{x})) \tag{3.2.6}$$

where we refer to L_1 and L_{-1} as the *partial losses* of L (Scott (2012)). 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 (3.2.6), the cost-sensitive optimization problem can then be formulated as:

$$\begin{aligned} f^* = \arg\min_{f \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} \left[\delta \mathbb{1}_{\{y_i=1\}} \log\left(1 + e^{-f(\mathbf{x}_i)}\right) + (1-\delta)\mathbb{1}_{\{y_i=-1\}} \log\left(1 + e^{f(\mathbf{x}_i)}\right) \right] \\ + \gamma_{\mathcal{H}} \|f\|^2_{\mathcal{H}} + \gamma_{\mathcal{M}} \mathbf{f^T L f} \end{aligned} \tag{3.2.7}$$

We refer to the optimization problem in (3.2.7) as *Cost-sensitive Laplacian Kernel Logistic Regression* (Cos-LapKLR). The extensions of standard regularization algorithms by solving the optimization problems (posed in (3.2.1)) for different choices of cost function L and regularization parameters $\gamma_{\mathcal{H}}$ and $\gamma_{\mathcal{M}}$ have been developed (Belkin et al. (2006)). 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 (3.2.7) 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 (3.2.7) 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, (3.2.7) can then be equivalently re-written in a finite dimensional form as:

$$\begin{aligned} H(\boldsymbol{\alpha}) = \min_{\boldsymbol{\alpha} \in \mathbb{R}^{l+u}} \frac{1}{2l} \Big[\delta \mathbf{1} \left(\mathbf{1} + \mathbf{y}\right)^T \log\left(\mathbf{1} + e^{-(\mathbf{K}_L \mathbf{P}_L \boldsymbol{\alpha})}\right) + \\ + (1-\delta)\mathbf{1} \left(\mathbf{1} - \mathbf{y}\right)^T \log\left(\mathbf{1} + e^{(\mathbf{K}_L \mathbf{P}_L \boldsymbol{\alpha})}\right) \Big] + \gamma_{\mathcal{H}} \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} + \gamma_{\mathcal{M}} \boldsymbol{\alpha}^T \mathbf{K L K} \boldsymbol{\alpha} \end{aligned} \tag{3.2.8}$$

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 (3.2.8) 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 (Dennis and Moré (1977)). We used the limited-memory BFGS (LM-BFGS), which is an extension to the BFGS algorithm which uses a limited amount of computer memory (Byrd et al. (1995)).

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 $f:\mathbb{R}^{(l+u)}\rightarrow\mathbb{R}$

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 $\mathbf{D}=\mathrm{diag}\,(d_1,\ldots,d_{l+u})$ and $d_i=\sum_{j=1}^{l+u} w_{ij}$.

Step 4: Choose the regularization parameters $\gamma_{\mathcal{H}}$ and $\gamma_{\mathcal{M}}$, and the cost parameter δ.

Step 5: Compute $\boldsymbol{\alpha}^*$ using (3.2.8) 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) (Breiman (2001)), SVM (Vapnik (2013)), and AdaBoost (Friedman et al. (2000)). 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 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 (Veropoulos et al. (1999)). The regularization parameter C in cost-sensitive SVM (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 ROS, 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 (Chawla et al. (2002)). 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.

3.2.1. 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 Hsu et al. (2003) recommended for SVM, the values of the remaining parameters 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 (Belkin et al. (2006)):

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

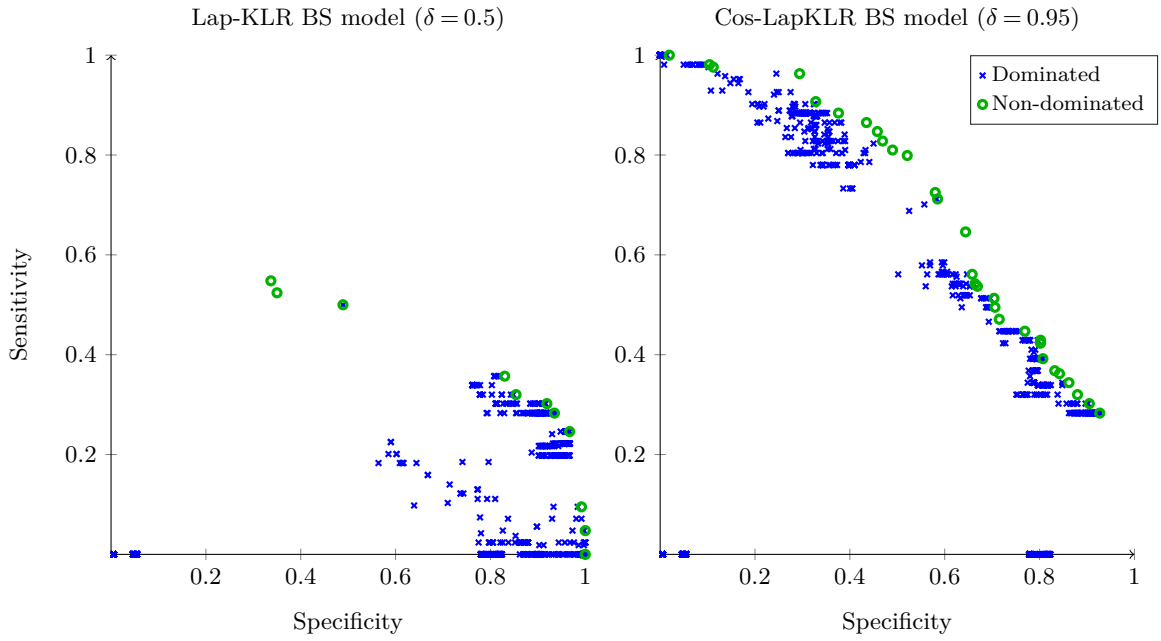
We applied the Pareto frontier based approach to select the optimal classifiers for each of these methods for distinguishing patients with metastasis at different cost setups during the training process.

For RF, we used the nominal values recommended by Friedman et al. (2001) 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 (Friedman et al. (2000), Schapire (2003)). 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. 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 3 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.

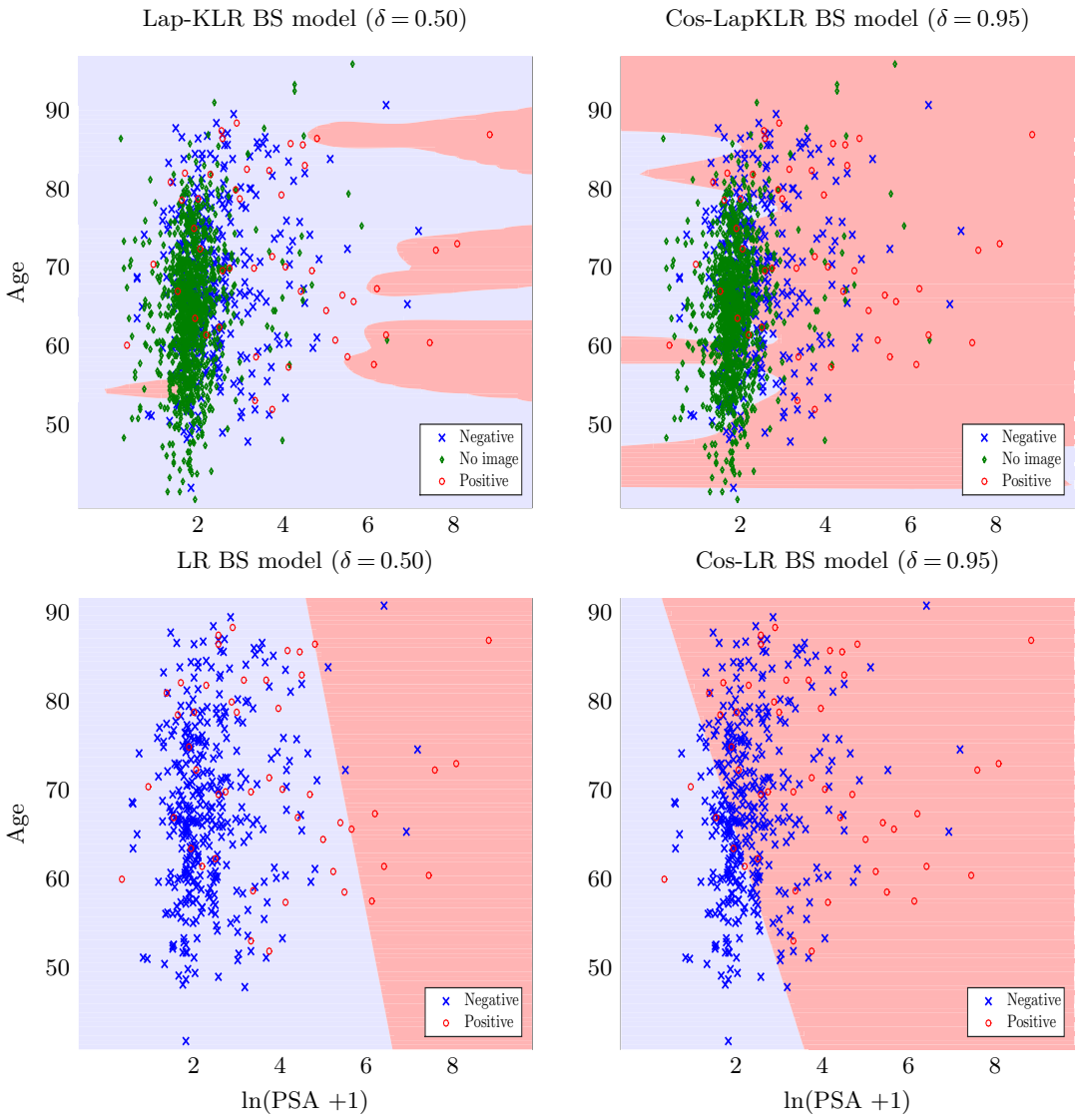
Figure 3 Pareto frontier graphs demonstrating the efficient frontiers based on sensitivity and specificity for Laplacian models predicting BS outcomes.



Our next set of experiments, in Figure 4, 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 ($\ln(\text{PSA}+1)$ 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, sensitivity of the classification rule increases while specificity decreases with increasing values of δ.

Figure 4 The impact of unequal misclassification costs on the decision boundaries of Cos-LR and CosLap-KLR.



4. Bias-corrected Performance of Imaging Guidelines

The results presented in Section 3.2.1 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.

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 are 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*) (Begg (1987)). This can markedly increase the apparent sensitivity of the test and reduce its apparent specificity (Begg (1987), Pepe (2003), Kosinski and Barnhart (2003)).

Several approaches have been proposed to address the problem of verification bias (Zhou (1998), Zhou et al. (2009)). 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 (Little (1988)). 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 (Begg and Greenes (1983)). 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 prostate cancer 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 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.

4.2. Bias-Corrected Results

There are several published clinical guidelines for BS and CT scans based on patient prostate cancer characteristics. These guidelines are summarized in Table 3. Table 4 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 Urological Association (AUA) guideline (Thompson et al. (2007)) 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 3 Published clinical guidelines for recommending BS and CT scan.

Bone scan		**CT scan**	
Clinical guidelines	**Recommend imaging if any of these:**	**Clinical guidelines**	**Recommend imaging if any of these:**
EAU (Mottet et al. (2014))	GS ≥ 8 cT3/T4 disease PSA > 10 ng/ml Symptomatic	**EAU** (Heidenreich et al. (2014))	GS ≥ 8 cT3/T4 disease PSA > 10 ng/ml Symptomatic
AUA (Thompson et al. (2007))	GS ≥ 8 PSA > 10 ng/ml Symptomatic	**AUA** (Carroll et al. (2013))	GS ≥ 8 PSA > 20 ng/ml cT3/T4 disease Symptomatic
NCCN (NCCN (2014))	cT1 disease & PSA > 20 ng/ml cT2 disease & PSA > 10 ng/ml GS ≥ 8 cT3/T4 disease Symptomatic		
Briganti's CART (Briganti et al. (2010))	GS ≥ 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.

We applied the bias-correction method on the optimized classification models of Section 3. Figure 5 shows the Pareto frontier graph consisting of all the imaging guidelines. The results indicate that the classification rules obtained using the methods of Section 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

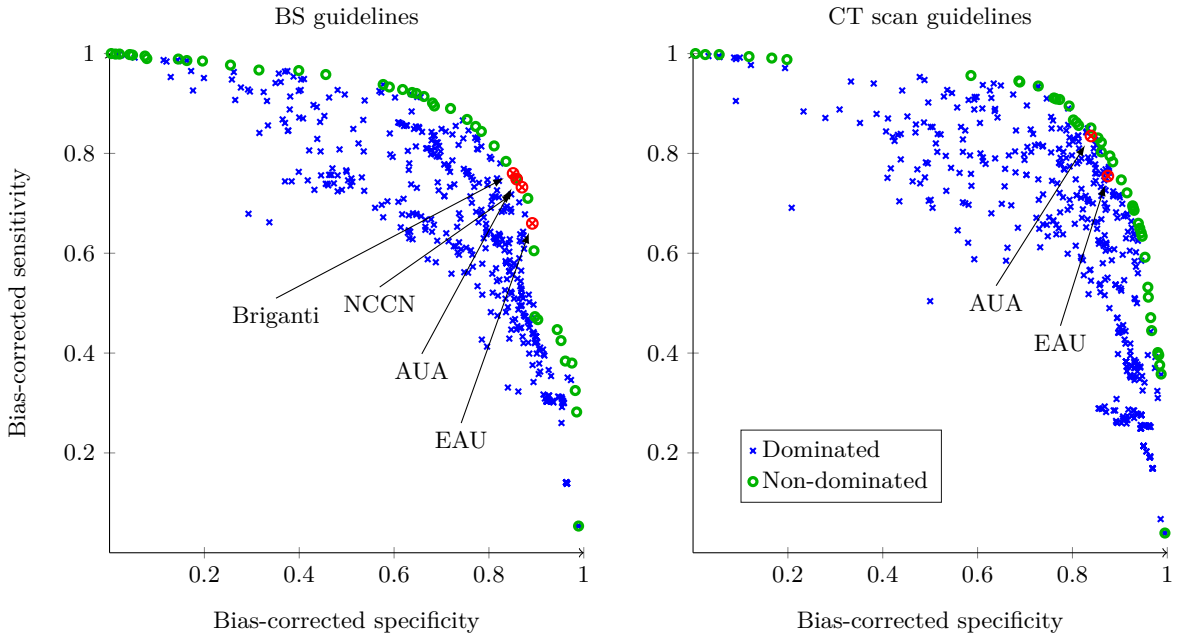
Table 4 Performance characteristics of the published guidelines before and after correcting for verification bias.

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

The numbers are the percentages.

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 3 but were all close to the Pareto frontier.

Figure 5 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 5 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 5 Proportions of classification modeling techniques that are non-dominated with respect to the bias-corrected accuracy.

Statistical models	**Bone scan (n = 40)**	**CT scan (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.

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 prostate cancer.

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:

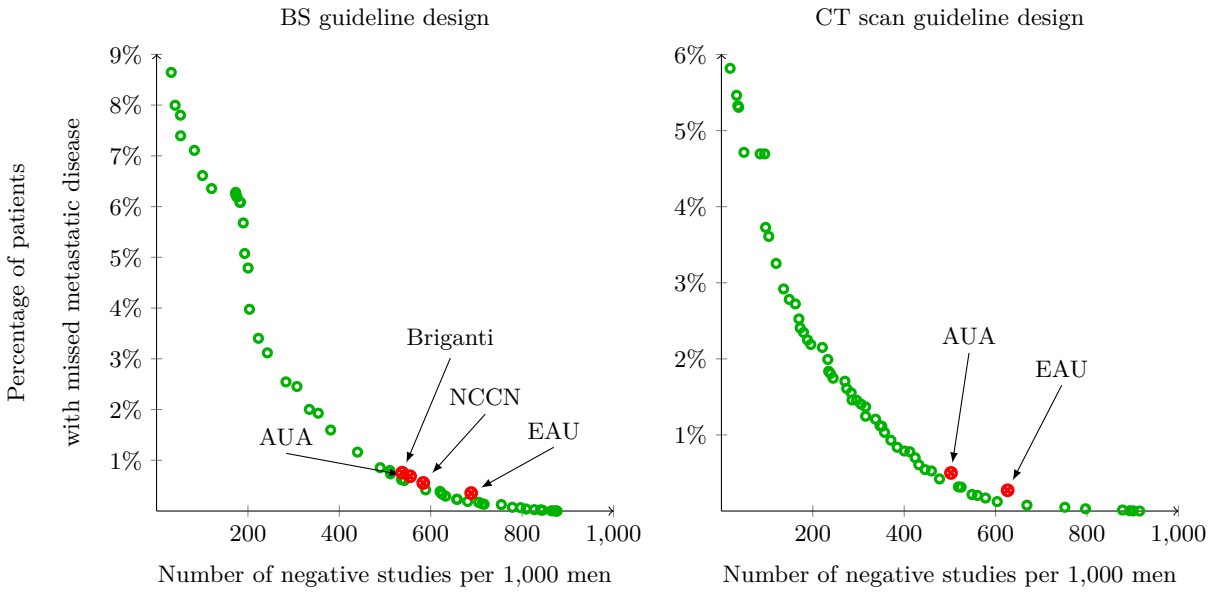
$$\mathbb{E}[Z^+] = \sum_{i=1}^{n} p_i \left(1 - g_i\right), \quad \mathbb{E}[Z^-] = \sum_{i=1}^{n} \left(1 - p_i\right) g_i$$

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 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 6 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 6 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.



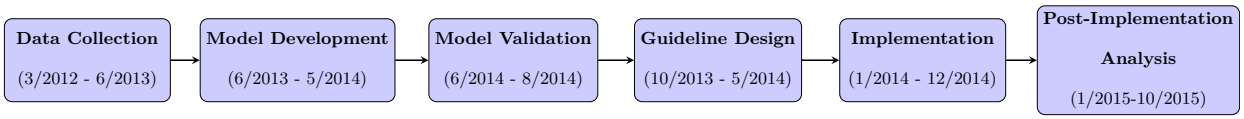
Additionally, we estimated the change in total number of imaging tests that can be expected from successful implementation of each clinical guideline compared to current practice. After assessing the performance of the available clinical guidelines on the appropriate use of BS and CT scan in newly-diagnosed prostate cancer 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% (Merdan et al. (2014), Risko et al. (2014)).

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 7. 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.

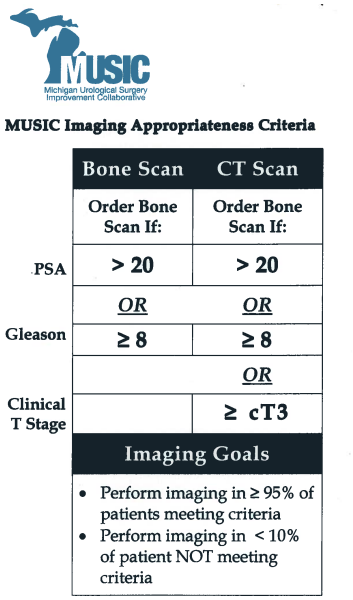
Figure 7 Project timeline from data collection to post-implementation analysis.



Our results and the resulting proposed guidelines were first reviewed by the MUSIC 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 ≥ 8 and recommends a CT scan for patients with PSA > 20 ng/mL, Gleason score ≥ 8, or clinical T stage $\geq$ cT3.

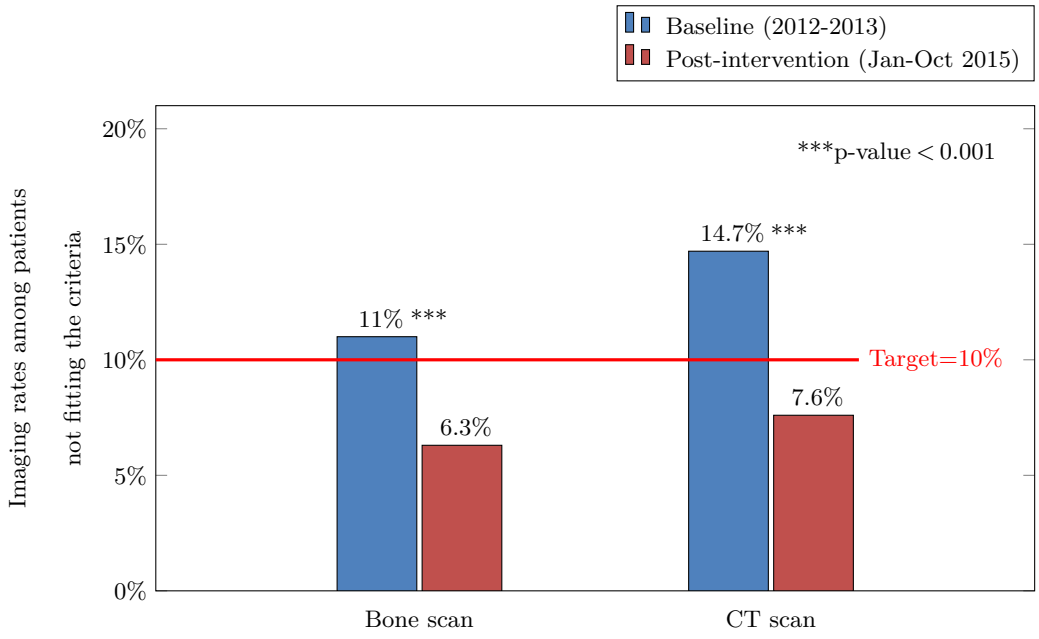
Figure 8 Placard sent to all urologists in the 43 MUSIC practices illustrating the selected imaging guidelines to be implemented.



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 8) 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.

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 9, 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 (Hurley et al. (2016)).

Figure 9 Avoidance of low-value imaging using MUSIC Criteria.



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. 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 prostate cancer patients was raised as the top priority by the American Urological Association "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 (Merdan et al. (2014), Risko et al. (2014), Hurley et al. (2016)). Recently, our publications have been cited in the new NCCN guidelines (NCCN (2014)). 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/`).

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. CMMI-1536444. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors would like to thank Susan Linsell and the MUSIC collaborative for their aide in this project.

Appendix A: 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 of the minority class (ROS) and random undersampling of the majority class (RUS).

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 (Chawla et al. (2002)). 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 (Han et al. (2005)). In order to obtain well-defined class clusters, several data cleaning methods such as the Edited Nearest Neighbor (ENN) rule (Batista et al. (2004)) and Tomek links (Tomek (1976)) have been integrated with SMOTE. SMOTE combined with two data cleaning techniques, Tomek links and ENN Rule (Wilson (1972)), 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) (Hart (1968)) and Neighborhood Cleaning Rule (NCR) (Laurikkala (2001)). 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 (Lemaître et al. (2017)). We performed 10 independent runs of 2-fold cross validation 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 BS and CT scan data sets developed by 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 AUC in both BS and CT scan datasets. These results clearly illustrate the inadequacy of the baseline RF and AdaBoost in recognizing metastatic patients.

References

Batista GE, Prati RC, Monard MC (2004) A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter* 6(1):20–29.

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

	BS (n = 416)				**CT scan (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.

Begg CB (1987) Biases in the assessment of diagnostic tests. *Statistics in Medicine* 6(4):411–423.

Begg CB, Greenes RA (1983) Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* 207–215.

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

Bleeker S, Moll H, Steyerberg E, Donders A, Derksen-Lubsen G, Grobbee D, Moons K (2003) External validation is necessary in prediction research:: A clinical example. *Journal of Clinical Epidemiology* 56(9):826–832.

Breiman L (2001) Random forests. *Machine Learning* 45(1):5–32.

Briganti A, Passoni N, Ferrari M, Capitanio U, Suardi N, Gallina A, Da Pozzo LF, Picchio M, Di Girolamo V, Salonia A, et al. (2010) 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. *European Urology* 57(4):551–558.

Byrd RH, Lu P, Nocedal J, Zhu C (1995) A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* 16(5):1190–1208.

Carroll P, Greene K, Babaian RJ, H Ballentine Carter PHG, Han M, Kuban DA, Sartor AO, Stanford JL, Zietman A (2013) PSA testing for the pretreatment staging and posttreatment management of prostate cancer: 2013 revision of 2009 best practice statement. *American Urological Association* URL `https://www.auanet.org/guidelines/prostate-specific-antigen-(2009-amended-2013)`.

Chapelle O, Schlkopf B, Zien A (2010) *Semi-Supervised Learning* (The MIT Press), 1st edition.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16(1):321–357.

Chawla NV, Japkowicz N, Kotcz A (2004) Editorial: special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter* 6(1):1–6.

Cox DR (1958) Two further applications of a model for binary regression. *Biometrika* 562–565.

Cruz JA, Wishart DS (2006) Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics* 2:59.

Dennis JE Jr, Moré JJ (1977) Quasi-newton methods, motivation and theory. *SIAM review* 19(1):46–89.

Domingos P (1999) Metacost: A general method for making classifiers cost-sensitive. *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 155–164 (ACM).

Efron B, Tibshirani R (1997) Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association* 92(438):548–560.

Efron B, Tibshirani RJ (1994) *An introduction to the bootstrap* (CRC press).

Elkan C (2001) The foundations of cost-sensitive learning. *International Joint Conference on Artificial Intelligence*, volume 17, 973–978 (Citeseer).

Evgeniou T, Pontil M, Poggio T (2000) Regularization networks and support vector machines. *Advances in Computational Mathematics* 13(1):1–50.

Friedman J, Hastie T, Tibshirani R (2001) *The elements of statistical learning*, volume 1 (Springer).

Friedman J, Hastie T, Tibshirani R, et al. (2000) Additive logistic regression: a statistical view of boosting. *The Annals of Statistics* 28(2):337–407.

Greiner R, Grove AJ, Roth D (2002) Learning cost-sensitive active classifiers. *Artificial Intelligence* 139(2):137 – 174.

Han H, Wang WY, Mao BH (2005) Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. *Advances in Intelligent Computing* 878–887.

Harrell F, Lee KL, Mark DB (1996) Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 15:361–387.

Hart P (1968) The condensed nearest neighbor rule. *IEEE Transactions on Information Theory* 14(3):515–516.

He H, Garcia EA (2009) Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on* 21(9):1263–1284.

Heidenreich A, Bastian PJ, Bellmunt J, Bolla M, Joniau S, van der Kwast T, Mason M, Matveev V, Wiegel T, Zattoni F, Mottet N (2014) EAU guidelines on prostate cancer. Part II: Treatment of advanced, relapsing, and castration-resistant prostate cancer. *European Urology* 65(2):467 – 479, ISSN 0302-2838, URL `http://dx.doi.org/http://dx.doi.org/10.1016/j.eururo.2013.11.002`.

Hsu CW, Chang CC, Lin CJ (2003) A practical guide to support vector classification. Technical report, National Taiwan University, Department of Computer Science, URL `http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`.

Hurley P, Montie J, Dhir A, Gao Y, Drabik B, Lim K, Curry J, Linsell S, Brachulis A, Ghani K, Denton B, Miller D (2016) A statewide intervention to reduce the use of low value imaging among men with newly-diagnosed prostate cancer. *The Journal of Urology* 195(4):591–592.

Kimeldorf G, Wahba G (1971) Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications* 33(1):82–95.

Kosinski AS, Barnhart HX (2003) Accounting for nonignorable verification bias in assessment of diagnostic tests. *Biometrics* 59(1):163–171.

Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI (2015) Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal* 13:8–17.

Laurikkala J (2001) Improving identification of difficult small classes by balancing class distribution. *Conference on Artificial Intelligence in Medicine in Europe*, 63–66 (Springer).

Lemaître G, Nogueira F, Aridas CK (2017) Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research* 18(17):1–5, URL `http://jmlr.org/papers/v18/16-365.html`.

Li Y, Kwok JTY, Zhou ZH (2010) Cost-sensitive semi-supervised support vector machine. *Proceedings of the National Conference on Artificial Intelligence*, volume 1, 500.

Little RJ (1988) A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association* 83(404):1198–1202.

Liu A, Jun G, Ghosh J (2009) Spatially cost-sensitive active learning. *Proceedings of the 2009 SIAM International Conference on Data Mining*, 814–825 (SIAM).

Liu XY, Zhou ZH (2006) The influence of class imbalance on cost-sensitive learning: An empirical study. *Data Mining, 2006. ICDM'06. Sixth International Conference on*, 970–974 (IEEE).

Maalouf M, Trafalis TB, Adrianto I (2011) Kernel logistic regression using truncated newton method. *Computational Management Science* 8(4):415–428.

Maloof MA (2003) Learning when data sets are imbalanced and when costs are unequal and unknown. *ICML-2003 Workshop on Learning from Imbalanced Datasets II*, volume 2, 2–1.

Margineantu DD (2005) Active cost-sensitive learning. *IJCAI*, volume 5, 1622–1623.

Masnadi-Shirazi H, Vasconcelos N (2010) Risk minimization, probability elicitation, and cost-sensitive svms. *ICML*, 759–766.

McCarthy K, Zabar B, Weiss G (2005) Does cost-sensitive learning beat sampling for classifying rare classes? *Proceedings of the 1st International Workshop on Utility-based Data Mining*, 69–77 (ACM).

Merdan S, Womble PR, Miller DC, Barnett C, Ye Z, Linsell SM, Montie JE, Denton BT (2014) Toward better use of bone scans among men with early-stage prostate cancer. *Urology* 84(4):793–798.

Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ (1993) Validation of probabilistic predictions. *Medical Decision Making* 13(1):49–57.

Mottet N, Bellmunt J, Briers E, Association EU, et al. (2014) Guidelines on prostate cancer. *European Urology* 65(1):124–37.

NCCN (2014) National Comprehensive Cancer Network Clinical Guidelines. `https://www.nccn.org/professionals/physician_gls/f_guidelines.asp`.

Pepe MS (2003) *The statistical evaluation of medical tests for classification and prediction* (Oxford University Press).

Qi Z, Tian Y, Shi Y, Yu X (2013) Cost-sensitive support vector machine for semi-supervised learning. *Procedia Computer Science* 18:1684 – 1689, international Conference on Computational Science.

Qin Z, Zhang S, Liu L, Wang T (2008) Cost-sensitive semi-supervised classification using cs-em. *8th IEEE International Conference on Computer and Information Technology*, 131–136 (IEEE).

Risko R, Merdan S, Womble PR, Barnett C, Ye Z, Linsell SM, Montie JE, Miller DC, Denton BT (2014) Clinical predictors and recommendations for staging computed tomography scan among men with prostate cancer. *Urology* 84(6):1329–1334.

Schapire RE (2003) The boosting approach to machine learning: An overview. *Nonlinear Estimation and Classification*, 149–171 (Springer).

Scott C (2012) Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics* 6:958–992.

Shawe-Taylor J, Cristianini N (2004) *Kernel methods for pattern analysis* (Cambridge University Press).

Sindhwani V, Niyogi P, Belkin M, Keerthi S (2005) Linear manifold regularization for large scale semi-supervised learning. *Proc. of the 22nd ICML Workshop on Learning with Partially Classified Training Data*, volume 28.

Sun Y, Kamel MS, Wong AK, Wang Y (2007) Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition* 40(12):3358–3378.

Thompson I, Thrasher J, Aus G, Burnett A, Canby-Hagino E, Cookson M, D'Amico A, Dmochowski R, Eton D, Forman J, Goldenberg S, Hernandez J, Higano C, Kraus S, Moul J, Tangen C (2007) Guideline for

the management of clinically localized prostate cancer: 2007 update. *Journal of Urology* 177(6):2106–2131, ISSN 0022-5347, URL `http://dx.doi.org/10.1016/j.juro.2007.03.003`.

Ting KM (2002) An instance-weighting method to induce cost-sensitive trees. *IEEE Transactions on Knowledge and Data Engineering* 14(3):659–665.

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

Tomek I (1976) Two modifications of CNN. *IEEE Transactions on Systems, Man and Cybernetics* SMC-6(11):769–772.

Vapnik V (2013) *The nature of statistical learning theory* (Springer Science & Business Media).

Veropoulos K, Campbell C, Cristianini N (1999) Controlling the sensitivity of support vector machines. *Proceedings of the International Joint Conference on Artificial Intelligence*, 55–60.

Von Luxburg U (2007) A tutorial on spectral clustering. *Statistics and Computing* 17(4):395–416.

Weiss GM (2004) Mining with rarity: a unifying framework. *ACM Sigkdd Explorations Newsletter* 6(1):7–19.

Wilson DL (1972) Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics* 2(3):408–421.

Zhou XH (1998) Correcting for verification bias in studies of a diagnostic test's accuracy. *Statistical Methods in Medical Research* 7(4):337–353.

Zhou XH, McClish DK, Obuchowski NA (2009) *Statistical methods in diagnostic medicine*, volume 569 (John Wiley & Sons).

Zhou ZH, Li M (2010) Semi-supervised learning by disagreement. *Knowledge and Information Systems* 24(3):415–439.

Zhou ZH, Liu XY (2006) Training cost-sensitive neural networks with methods addressing the class imbalance problem. *Knowledge and Data Engineering, IEEE Transactions* 18(1):63–77.

Zhu J, Hastie T (2005) Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics* 14(1).

Zhu X (2007) Semi-supervised learning literature survey. Technical report, Computer Science, University of Wisconsin-Madison, URL `http://pages.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf`.

Zhu X, Goldberg AB (2009) Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 3(1):1–130.