informs
http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–21
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Data Analytics for Optimal Detection of Metastatic Prostate Cancer

Selin Merdan,[a] Christine L. Barnett,[a] Brian T. Denton,[a] James E. Montie,[b,c] David C. Miller[b,c]

[a] Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, Michigan 48109; [b] Department of Urology, University of Michigan, Ann Arbor, Michigan 48109; [c] Michigan Urological Surgery Improvement Collaborative, Ann Arbor, Michigan 48109

Contact: smerdan@umich.edu (SM); clbarnet@umich.edu (CLB); btdenton@umich.edu, https://orcid.org/0000-0002-6372-6066 (BTD); jmontie@umich.edu (JEM,); dcmiller@umich.edu (DCM)

Received: August 10, 2017
Accepted: February 14, 2020
Published Online in Articles in Advance: ■■ ■■, 2020

Subject Classifications: programming: nonlinear, applications, statistics, healthcare
Area of Review: Practice

https://doi.org/10.1287/opre.2020.2020

Abstract. We used data-analytics approaches to develop, calibrate, and validate predictive models, to help urologists in a large statewide 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 the detection of metastatic prostate cancer. The Michigan Urological Surgery Improvement Collaborative, a statewide 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 after implementation to confirm their impact on reducing unnecessary imaging across the state of Michigan.

Funding: This work was supported by the National Science Foundation [Grant CMMI-1536444].

Keywords: healthcare • prostate cancer: radiographic staging • semisupervised learning • class imbalance problem • cost-sensitive learning • verification bias

1. Introduction

Prostate cancer is the most commonly diagnosed cancer and the third leading cause of death by cancer in American men (Hricak et al. 2007). It was estimated that in 2019, 174,650 news cases of prostate cancer would be diagnosed and approximately 31,620 men would die of the disease in the United States (American Cancer Society 2019a). For each newly diagnosed cancer case, clinical staging will be performed, which is an assessment of how far the cancer has spread based on the results from biopsy, blood tests, and imaging (American Cancer Society 2019b). Prostate cancer 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 prostate cancer; therefore, bone metastases are one of the most important prognostic factors (Tombal and Lecouvet 2012). Bone metastases are associated with considerable morbidity (pain, reduced mobility, pathological fractures, and spinal cord and nerve compression), reduced survival (five-year survival is 3%), and also significant economic health implications, including the costs of systematic therapies, imaging, and hospital admissions (Pockett et al. 2010, Yong et al. 2014, Broder et al. 2015). The presence of lymph node metastasis is also an important prognostic factor, indicating great risk for progression to bone metastases and death (Fuchsjäger et al. 2008). As of today, metastatic prostate cancer is still considered incurable, but there are treatment options that can increase survival. Therefore, accurate staging is crucial for the clinical management of prostate cancer, from possible cure (in patients without metastases) to alleviating symptoms and improving quality of life (in patients with metastases).

Conventional imaging tests for prostate cancer staging include bone scans (BSs) and computed tomography (CT) scans for detection of bone and lymph node metastases, respectively. However, not all men with newly diagnosed prostate cancer are at the same risk of harboring metastatic cancer. This is an important consideration, because there are harms associated with both under- and overimaging. Underimaging results in patients' metastatic prostate cancer going undetected. In such cases, patients are subjected to treatments, such as radical prostatectomy (surgical removal of the prostate), that are unlikely to benefit them and can lead to serious side effects and negative health outcomes due to delays in chemotherapy (Lavery et al. 2011, Prasad et al. 2012, Kim et al. 2015). Overimaging causes potentially harmful radiation exposure (Prasad et al. 2004, Smith-Bindman et al. 2009, Lin 2010), 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.

There are several international evidence-based guidelines indicating the need for BSs and CT scans 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 BSs and CT scans for men newly diagnosed with prostate cancer (Thompson et al. 2007, Briganti et al. 2010, Carroll et al. 2013, Heidenreich et al. 2014, Mottet et al. 2014, National Comprehensive Cancer Network 2014). 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. Because a randomized trial is not possible for practical and ethical reasons for the imaging of prostate cancer staging, we proposed data-driven approaches and evaluated them 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 while 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.

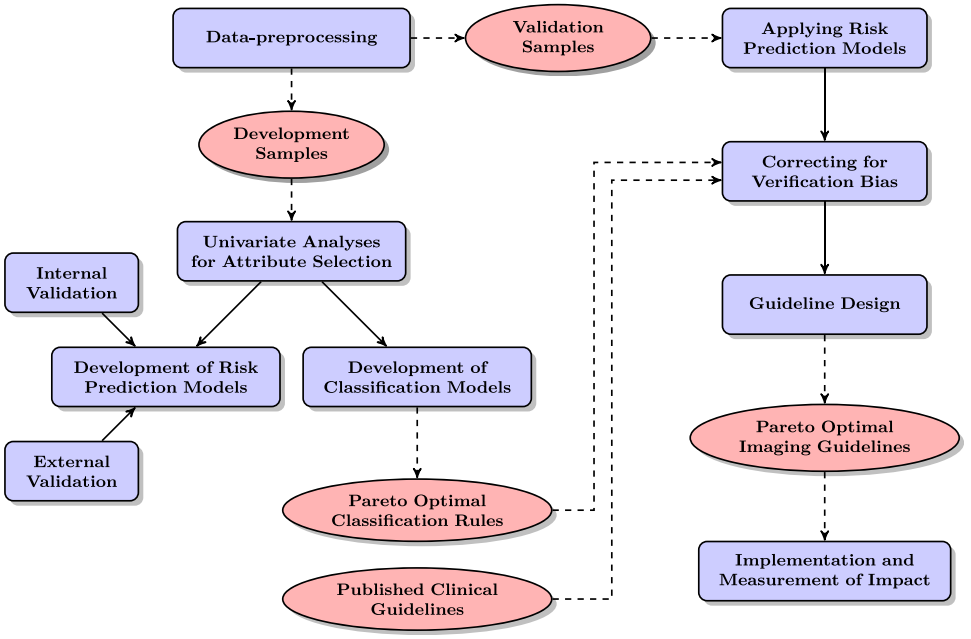
The most significant novel aspect of this study is the combination of multiple methods drawing from operations research, statistics, and machine learning to create a framework for addressing important decision-making problems in the context of imperfect observational data. This stems from the study's goal to make practical recommendations, which requires careful consideration of factors that are often overlooked in more methodologically focused studies. We are also, to the best of our knowledge, the first to simultaneously consider the exploitation of unlabeled data (for patients who did not have the gold-standard tests at diagnosis) and the incorporation of a cost-sensitive learning scheme to deal with class imbalance, based on a novel extension (discussed in Section 3) of the data-dependent geometric regularization framework proposed by Belkin et al. (2006).

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 paper is structured as follows. Section 2 describes the methodological approach for the 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 semisupervised 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 postimplementation analysis. Section 6 highlights our main conclusions and states some points for future research.

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

Figure 1. (Color online) Research Framework Illustrating the Major Steps from Data Preprocessing to Implementation and Measurement of Impact



Note. Straight lines represent the connections between major processing steps involved in the proposed research framework (in rectangles), and dashed lines represent the flow of input/output (in ellipses) between the processes.

models for estimating the probability of an imaging test that was positive for metastases.

2.1. Clinical Data Sets 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 a Gleason score (GS), a percentage of positive samples (also called *cores*) that show cancer, and the maximum percentage of core involvement. These risk factors are determined by review of biopsy samples by a trained pathologist. If cancer cells are found upon evaluation of biopsy samples, the pathologist will give the most common pattern a grade of 1 to 5. The second most common cell pattern will also be given a grade of 1 to 5. These two grades are then added together to obtain the patient's GS between 2 and 10 (Cancer.Net 2019). The 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 are 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. The 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. These three clinical risk factors (serum PSA at diagnosis, GS, and clinical T stage) help determine whether radiological imaging (BS or CT scan) is needed to complete the staging of the prostate cancer patient.

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. Appendix A presents the clinical characteristics of patients included in the analytic samples for BS and CT scan. 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. Patients who underwent imaging had significantly higher PSA levels, biopsy GS, and clinical T stages than those who did not receive imaging (Appendix A, all $p < 0.0001$).

We performed univariate and multivariate analyses to examine the association between imaging outcomes and all routinely available clinical variables in imaged patients. The results from these analyses are presented in our earlier work (Merdan et al. 2014, Risko et al. 2014). 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 because the distribution of PSA scores 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 use logistic regression (LR) models to estimate the probability of a positive imaging outcome. The discriminative model for LR is given by

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

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

$$\begin{aligned} -\log \mathbb{L}(\boldsymbol{\beta}) &= -\log \prod_{i=1}^{l} \mathbb{P}(y_i = \pm 1 \mid \mathbf{x}_i, \boldsymbol{\beta}) \\ &= \sum_{i=1}^{l} \log\left(1 + e^{-y_i \boldsymbol{\beta}^T \mathbf{x}_i}\right) \end{aligned} \tag{2}$$

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 data set to develop and validate the model, and external validation uses an independent data set 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 data set better than a new data set. Given the intention to implement these guidelines for clinical practice, it was necessary to carefully consider this bias. Although internal validation is commonly done by randomly splitting the data set into a training sample and a validation sample, this approach is argued to be statistically inefficient, as not all available data are used to develop the prediction model. Therefore, bootstrapping is the preferred method for internal validation, especially when the development sample is relatively small or a high number of candidate predictors is studied (Harrell et al. 1996, Efron and Tibshirani 1997). Given its efficiency in predictive modeling with logistic regression in small data sets, we used bootstrapping to obtain valid estimates of the expected optimism in predictive performance before we had enough out-of-sample data to validate our models. (Steyerberg et al. 2001, 2003).

Because 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 were 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 $\sum_{i=1}^{n}(y_i - \mathbb{P}(y_i = 1 \mid \mathbf{x}_i, \boldsymbol{\beta}))^2/n$, where we assume that n is the size of the sample with which the model is being assessed and $y \in \{0,1\}$. We introduced $y \in \{-1,1\}$ earlier in Section 2.2, as this labeling renders the formulation of the logistic loss function as partial losses introduced later in Section 3.2 easier to understand. 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 (LP) of the LR model, known as the *calibration slope* (Miller et al. 1993). The LP is the sum of the regression coefficients multiplied by the patient value of the corresponding predictor (i.e., for patient i, $\mathrm{LP}_i = \mathbf{x}_i\boldsymbol{\beta}$). 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, $\mathrm{logit}(\mathbb{P}(\mathrm{y} = 1)) = \alpha + \beta_{calibration}\mathrm{LP}$; 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 data set 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 (Miller et al. 1993, Harrell et al. 1996).

We assessed the model calibration graphically with calibration plots. We divided the patients into 10 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. It 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 the 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), where $P(\cdot)$ represents any of the three performance metrics that we described in the previous section. We used this approach to internally validate the model calibration and discrimination.

Algorithm 1 (Bootstrapping Algorithm for Internal Validation) Q: 21

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 data sets 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.

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 that for CT scan included 507 patients, of which 42 (8.28%) were interpreted as positive for lymph node metastasis. The change in the AUC between the internal and external validations 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.

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 significant 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 diagnosis, which results in a missing data problem. The second challenge, imbalanced data, arises from the fact that a minority of patients have metastatic cancer. To address each of these challenges, we study two machine

Table 1. Bootstrap Results for the Development Samples

	Development samples	
	BS (n = 416) mean ± $SE_{bootstrap}$	CT scan (n = 643) mean ± $SE_{bootstrap}$
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

Note. In the development samples for BS and CT scan, 1,000 bootstrap repetitions were used for the calculation of both the means and standard errors ($SE_{bootstrap}$).

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

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

learning paradigms in this article: *semisupervised* and *cost-sensitive* learning.

Semisupervised learning aims to improve the learning performance by appropriately exploiting the unlabeled data in addition to the labeled data (Zhu 2007, Zhu and Goldberg 2009, Chapelle et al. 2010, 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, whereas 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 (Domingos 1999, Elkan 2001, Ting 2002, Maloof 2003, 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).

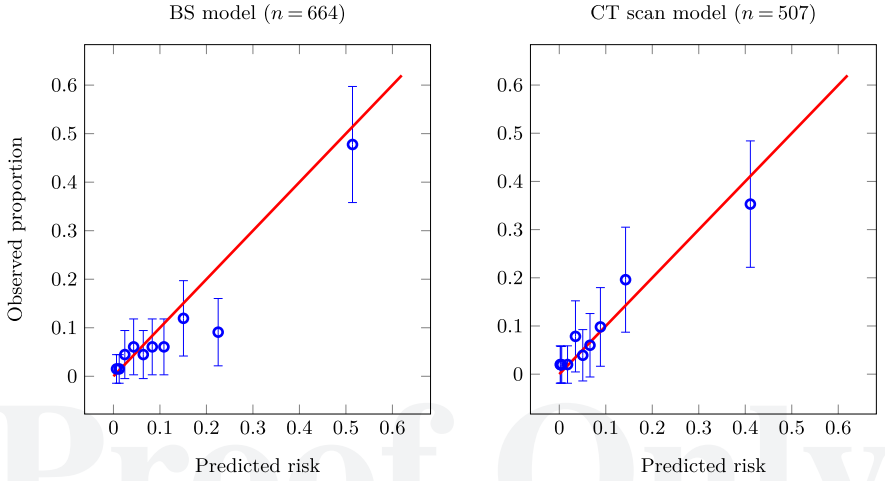
The use of unlabeled data in cost-sensitive learning has attracted growing attention, and many techniques have been developed (Greiner et al. 2002, Margineantu 2005, Qin et al. 2008, Liu et al. 2009, Li et al. 2010, Qi et al. 2013). To our knowledge, however, there has not been an attempt to apply both semisupervised and cost-sensitive learning to improve cancer diagnosis (see the literature reviews in Cruz and Wishart 2006, Kourou et al. 2015). 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 *overfitting* to the training data, which is widely used in machine learning (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 $\text{sign}(0) = -1$. Given a set of labeled data, Q: 23

Q: 72 **Figure 2.** (Color online) Calibration Plots for BS and CT Scan Risk Prediction Models Based on the Validation Samples



$\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)\} \in (\mathbb{R}^n \times \{-1, 1\})^l$, a general class of regularization problems estimates the unknown function f by minimizing the functional

$$\min_{f \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} L(y_i, f(\mathbf{x}_i)) + \gamma_{\mathcal{H}} \|f\|_{\mathcal{H}}^2, \tag{3}$$

where $L(y, f(\mathbf{x}))$ is the loss function, and $\|\cdot\|_{\mathcal{H}}$ 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, $\mathbf{K} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. 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) has the form

$$f^*(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i^* \mathbf{K}(\mathbf{x}, \mathbf{x}_i). \tag{4}$$

As a consequence, (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 (SVMs) and regularized least squares (Belkin et al. 2006).

In order to address the issue of missing data for patients who did not receive a BS or CT scan, we use the Laplacian semisupervised framework proposed by Belkin et al. (2006), which extends the classical framework of regularization given in (3) via manifold regularization. This framework relies on two basic assumptions: the *cluster assumption* and the *manifold assumption*. In our context, the former assumes that patients with similar characteristics should have similar observed imaging outcomes; the latter assumes that patients with similar characteristics should have similar predicted imaging outcomes. We extend their work by formulating the *logistic loss* for KLR, given as $L(y, f(\mathbf{x})) = \ln(1 + e^{-yf(\mathbf{x})})$, in terms of partial losses to adjust for class imbalance such that the cost of misclassifying a patient with metastasis outweighs the cost of misclassifying a cancer-free patient. In particular, we consider the following cost-sensitive optimization problem:

$$f^* = \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} \left[\delta L_1\big(f(\mathbf{x}_i)\big) + (1 - \delta) L_{-1}\big(f(\mathbf{x}_i)\big)\right] + \gamma_{\mathcal{H}} \|f\|_{\mathcal{H}}^2 + \gamma_{\mathcal{M}} \|f\|_{\mathcal{M}}^2, \tag{5}$$

where the partial losses L_1 and L_{-1} are defined as $L_1(f(\mathbf{x})) = \log(1 + e^{-f(\mathbf{x})})$ and $L_{-1}(f(\mathbf{x})) = \log(1 + e^{f(\mathbf{x})})$. The regularization parameters $\gamma_{\mathcal{H}}$ and $\gamma_{\mathcal{M}}$ control the $\mathcal{H}$ *norm* and the *intrinsic norm*, respectively. The first regularization term prefers the decision function to be a simple classifier, whereas the second term enforces that similar patients have similar imaging outcomes. We refer to the optimization problem in (5) as *cost-sensitive Laplacian kernel logistic regression* (*Cos-LapKLR*). More technical details about the manifold regularization in Cos-LapKLR and the proposed algorithmic solution are given in Appendix B. We also include the supervised cost-sensitive KLR, which we refer to as Cos-KLR, in our analyses.

In addition to Cos-LapKLR and Cos-KLR, we implemented and tested several other well-known classification models, including random forests (RFs; Breiman 2001), SVM (Vapnik 2013), and AdaBoost (Friedman et al. 2000). Scaling of (2) by a factor of $1/l$ establishes the equivalence between LR estimated by maximum likelihood and empirical risk minimization with logistic loss, where $f(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$, and $\boldsymbol{\beta} \in \mathbb{R}^d$ is a d-dimensional vector of patient attributes. 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.

Similar to Cos-LapKLR and Cos-KLR, 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 the cost-sensitive SVM (Cos-SVM) corresponds to the misclassification cost, which involves two parts, that is, the cost of misclassifying the negative class into the positive class and the cost of misclassifying the positive class into the negative class. In this work, the cost of misclassifying the negative class as positive is set to C, whereas the cost of misclassifying the positive class into the negative class is set to $C \times \delta/(1 - \delta)$, where $\delta \in (0, 1)$.

To remedy the class imbalance problem with RFs and AdaBoost, different data sampling techniques were employed in the experimental evaluation, such as random oversampling of the minority class (ROS), random undersampling of the majority class (RUS), and the combination of both methods. ROS and RUS are nonheuristic methods that were 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 C.

3.2.1. Classification Model Results. We adopted two-fold cross-validation (CV) in the model training process. The radial basis function kernel of the form $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$ 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 an HP Z230 work station with an Intel Xeon E31245W (3.4 GHz) processor with cores and 16 GB of random access memory. 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 the true positive rate, indicates the accuracy on the positive class; specificity, or the 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 nondominated models for varying choices of cost parameter based on twofold 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 that of 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 twofold CV at different cost setups. For Cos-LapKLR, candidate values for the regularization parameters $\gamma_{\mathcal{H}}$ and $\gamma_{\mathcal{M}}$ were chosen from the set $\{2^i \mid -13, -11, \ldots, 3\}$, the those for the kernel parameter γ from $\{2^i \mid -9, -7, \ldots, 3\}$, and the nearest neighbor parameter k from $\{3, 5\}$ (see Appendix B for the construction of the similarity matrix). For Cos-KLR, same ranges were used for $\gamma_{\mathcal{H}}$ and γ. For Cos-LR, candidate values for the regularization parameter λ were chosen from the set $\{2^i \mid -13, -11, \ldots, 3\}$. For Cos-SVM, candidate values for the regularization parameter C were chosen from the set $\{2^i \mid -5, -3, \ldots, 15\}$, and those for the kernel parameter γ from $\{2^i \mid -15, -13, \ldots, 3\}$. 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 RFs, 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, because this was shown to yield good performance of AdaBoost (Friedman et al. 2000, Schapire 2003). We performed 10 independent runs of twofold 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 C. In the remainder of this section, we summarize results for the cost-sensitive methods (i.e., Cos-LapKLR, Cos-KLR, 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-KLR, 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 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(PSA + 1) and age). We see that higher penalties on L_1 loss increase the region of $\mathbb{P}(y = 1 \mid \mathbf{x})$ (shaded area), corresponding to patients with predicted outcome $\hat{y} = 1$, that is, $f(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta} \geq 0$, and, thus, sensitivity of the classification rule increases and specificity decreases with increasing values of δ.

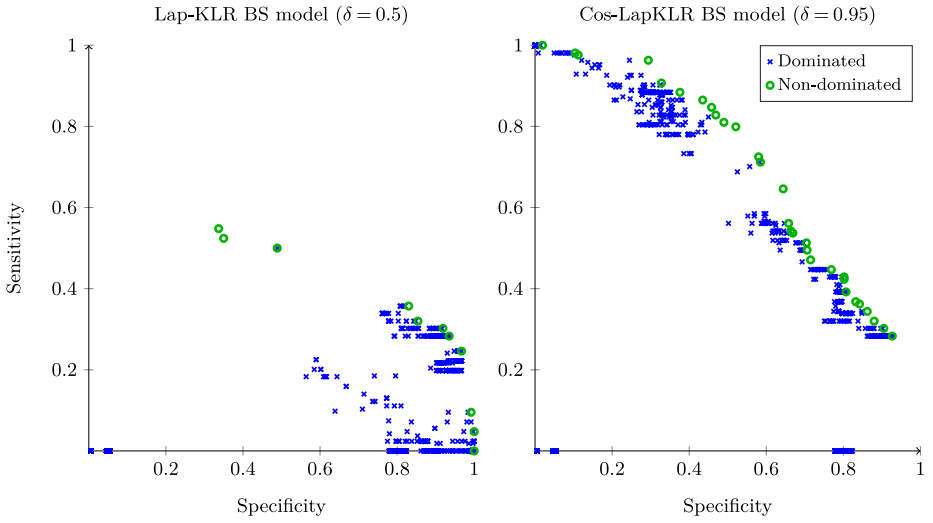
4. Bias-Corrected Performance of Imaging Guidelines

The results presented in Section 3.2.1 for the sensitivity and specificity of alternative classification models are systemically biased because 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

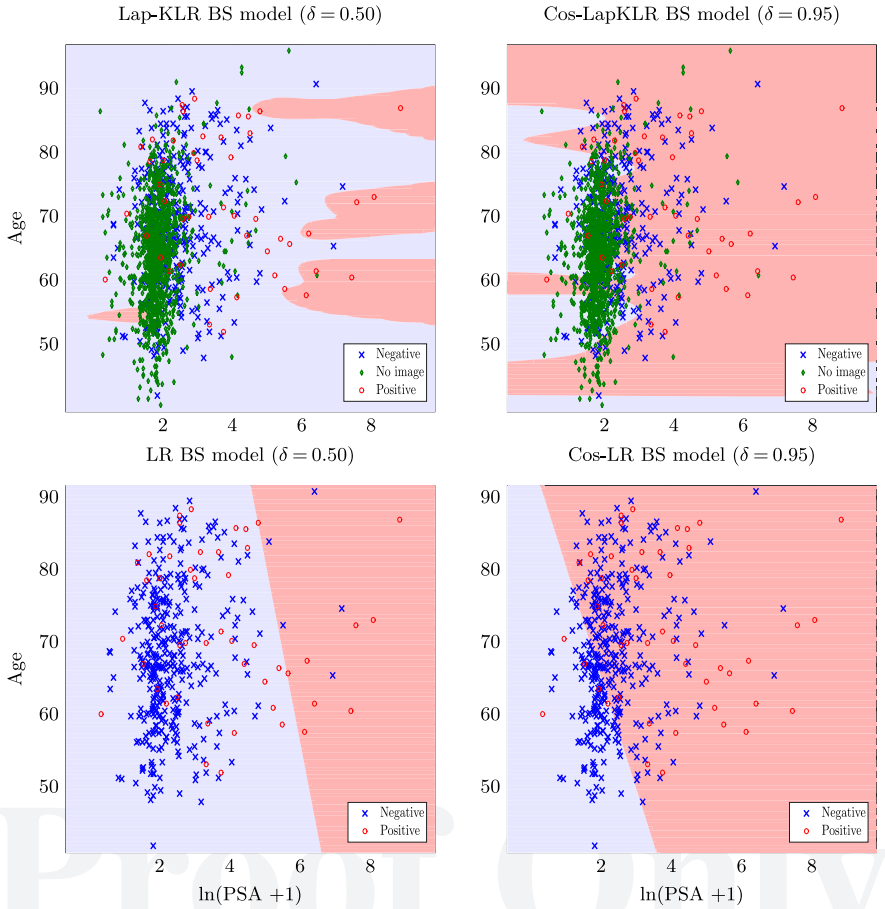
Standard inferential procedures rely on several assumptions concerning study design, such as the existence of a reference test, usually referred to as a gold standard, a procedure that is known to be capable of classifying an individual as diseased or nondiseased. In practice, gold-standard tests are often invasive and may be expensive (e.g., BSs and CT scans 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 the presence of the disease with a gold-standard test is often influenced by individual patient risk factors. Patients who appear to be at high

Figure 3. (Color online) Pareto Frontier Graphs Demonstrating the Efficient Frontiers Based on Sensitivity and Specificity for Laplacian Models Predicting BS Outcomes



risk of disease may be very likely to be offered a 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

Figure 4. (Color online) The Impact of Unequal Misclassification Costs on the Decision Boundaries of Cos-LR and CosLap-KLR is Illustrated for $\delta = 0.50$ vs. $\delta = 0.95$



apparent specificity (Begg 1987, Kosinski and Barnhart 2003, Pepe 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 their 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 does not depend on the observed or missing data.

Q: 31 To obtain unbiased estimates of sensitivity and specificity, Begg and Greenes (1983) developed a Q: 32 method based on MLE. This method uses the observed proportions of patients with and without the disease among the verified patients to calculate the expected proportions among nonverified patients. The proportions 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 are the only factors that influenced the 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 "+" or "−" denotes 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

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

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

$$\begin{aligned}\text{Sensitivity} &= \mathbb{P}(\text{G+} \mid \text{Disease present}) \\ &= \frac{\mathbb{P}(\text{Disease present} \mid \text{G+})\mathbb{P}(\text{G+})}{\mathbb{P}(\text{Disease present})},\\ \text{Specificity} &= \mathbb{P}(\text{G−} \mid \text{Disease not present}) \\ &= \frac{\mathbb{P}(\text{Disease not present} \mid \text{G−})\mathbb{P}(\text{G−})}{\mathbb{P}(\text{Disease not present})},\end{aligned}$$

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

$$\begin{aligned}\mathbb{P}(\text{Disease present}) &= \mathbb{P}(\text{Disease present} \mid \text{G+}) \\ &\quad \times \mathbb{P}(\text{G+}) + \mathbb{P}(\text{Disease present} \mid \text{G−})\mathbb{P}(\text{G−}),\\ \mathbb{P}(\text{Disease not present}) &= \mathbb{P}(\text{Disease not present} \mid \text{G+}) \\ &\quad \times \mathbb{P}(\text{G+}) + \mathbb{P}(\text{Disease not present} \mid \text{G−})\mathbb{P}(\text{G−})\end{aligned}.$$

Thus, to estimate the sensitivity and specificity of each guideline, we need to calculate $\mathbb{P}$(Disease present | G+), $\mathbb{P}$(Disease not present | G−), $\mathbb{P}$(G+), and $\mathbb{P}$(G−). To estimate $\mathbb{P}$(Disease present | G+) and $\mathbb{P}$(Disease not present | 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 $\mathbb{P}$(Disease present | 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 $\mathbb{P}$(Disease not present | 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 $\mathbb{P}$(G+) Q: 33 and $\mathbb{P}$(G−) as the proportions of the population in G+ and G−, respectively. 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 BSs 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, whereas 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.

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

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 > 20 ng/mL Symptomatic	AUA (Carroll et al. 2013)	GS ≥ 8 PSA > 20 ng/mL cT3/T4 disease Symptomatic
NCCN (National Comprehensive Cancer Network 2014)	cT1 disease and PSA > 20 ng/mL cT2 disease and PSA > 10 ng/mL GS ≥ 8 cT3/T4 disease Symptomatic		
Briganti's CART (Briganti et al. 2010)	GS ≥ 8 ≥ cT2 disease and PSA > 10 ng/mL Symptomatic		

Note. EAU, European Urological Association; CART, classification and regression tree.

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 3 but were all close to the Pareto frontier.

To further assess the performance of the statistical methods, we determine the proportions of the nondominated 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 estimated number of negative imaging tests. This finding underscores the importance of employing multiple methods for optimization of classification rules.

4.3. Patient-Centered Criteria

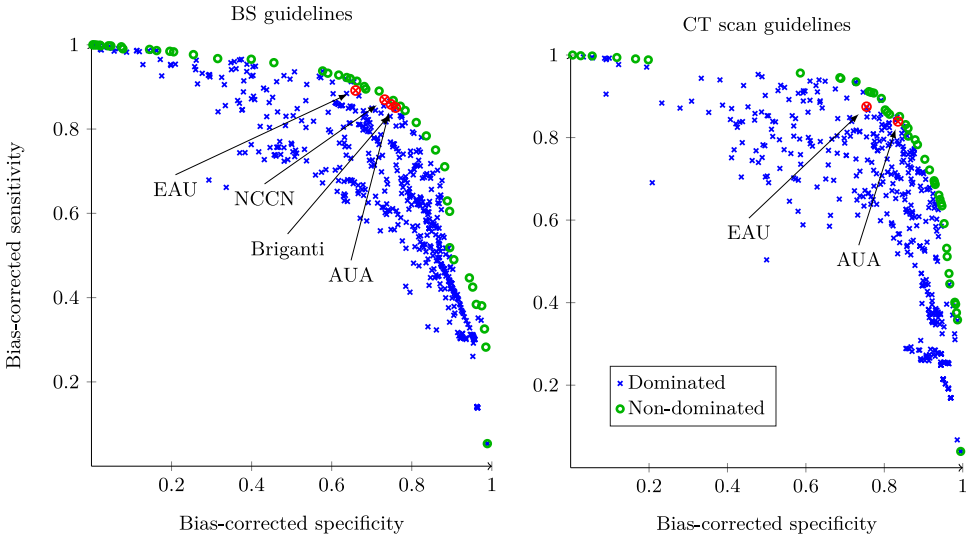
In working with MUSIC, we found that interpreting Q: 34 the results was easier when they were presented in terms of more patient-centered health outcomes. Therefore, we considered two important criteria: expected number of positive outcomes missed and expected number of negative studies. These estimates around the impact of specific guideline implementation can

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
Q: 75 Briganti's CART (Briganti et al. 2010)	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

Notes. The numbers are percentages. EAU, European Urological Association; CART, classification and regression tree.

Figure 5. (Color online) 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



Q: 73 *Note.* EAU, European Urological Association; Briganti, Briganti et al. (2010) method.

provide useful information for clinicians, specialty societies, and other stakeholders seeking a satisfactory trade-off between the benefits and harms of using these imaging tests for the staging of patients with newly diagnosed prostate cancer.

Q: 35 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 let it be 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

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

Table 5. Proportions of Classification Modeling Techniques That Are Nondominated with Respect to the Bias-Corrected Accuracy

Statistical models	Bone scan ($n = 44$)	CT scan ($n = 42$)
Cos-LapKLR	2.27	2.38
Cos-KLR	13.64	0.00
Cos-LR	43.18	64.29
Cos-SVM	25.00	21.43
RFs	15.91	11.90
AdaBoost	0.00	0.00

Q: 76 *Note.* The numbers are percentages.

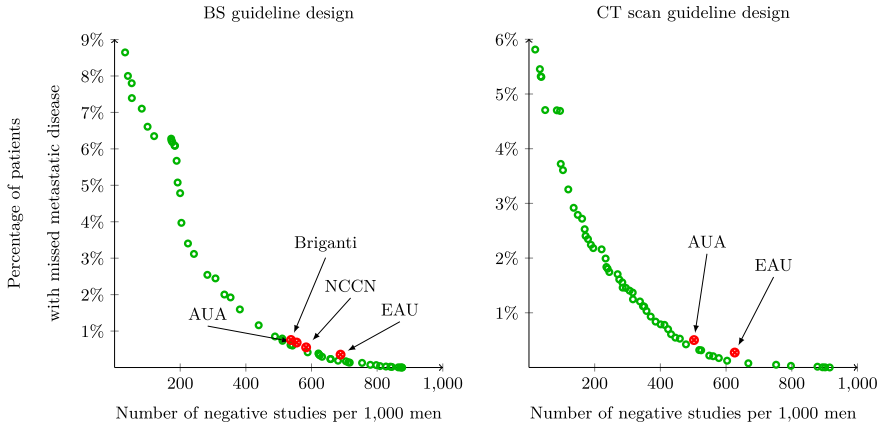
where $\mathbb{E}$ is the expectation operator. Assuming the goal is to find an optimal guideline that minimizes an *unweighted* function of these two competing criteria, the optimization model can be expressed as

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

where G is the set of all imaging guidelines consisting of the published clinical guidelines and the nondominated 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 of <1%.

Additionally, we estimated the change in total number of imaging tests that can be expected from successful implementation of each clinical guideline compared with 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 numbers of BS and CT scans by 25% and 26%, respectively, compared with 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). Based on the discussions with urologists, other clinicians, and patient advocates, 1% was deemed an acceptable miss rate in light of the significant reduction in

Figure 6. (Color online) 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



Note. EAU, European Urological Association; Briganti, Briganti et al. (2010) method.

unnecessary imaging. The AUA guideline achieved this miss rate and was very close to Pareto optimal. Moreover, the AUA guidelines carried the endorsement of a professional medical society. For these recommendations, we proposed the AUA guideline as the final criteria to be implemented, with the support of our data-driven models.

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.

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 BSs and CT scans, known as the MUSIC Imaging Appropriateness Criteria (see the YouTube video at https://youtu.be/FEIxb_HRHAA). The criteria recommend a BS for patients with a PSA score of >20 ng/mL or Gleason score ≥8, and recommend a CT scan for patients with a PSA score of >20 ng/mL, Gleason score ≥8, or clinical T stage ≥cT3.

Recognizing the importance of clinical judgment in staging decisions, the MUSIC consortium set a statewide goal of performing imaging in ≥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,

Figure 7. (Color online) Project Timeline from Data Collection to Postimplementation Analysis

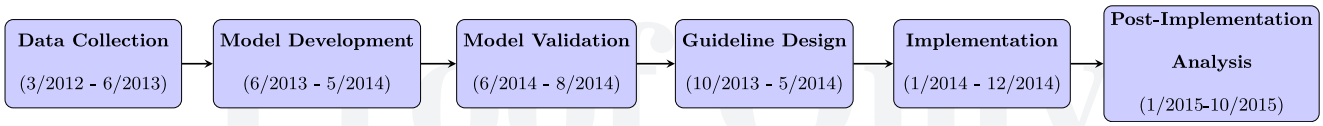
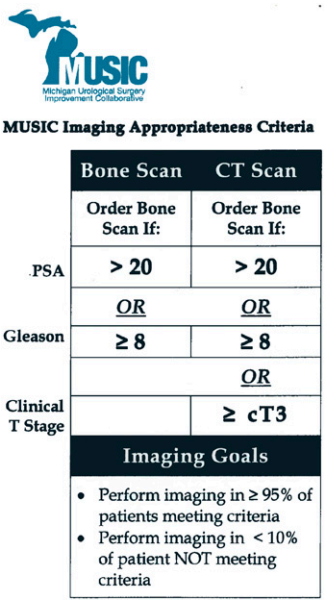


Figure 8. (Color online) Placard Sent to All Urologists in the 43 MUSIC Practices Illustrating the Selected Imaging Guidelines to Be Implemented



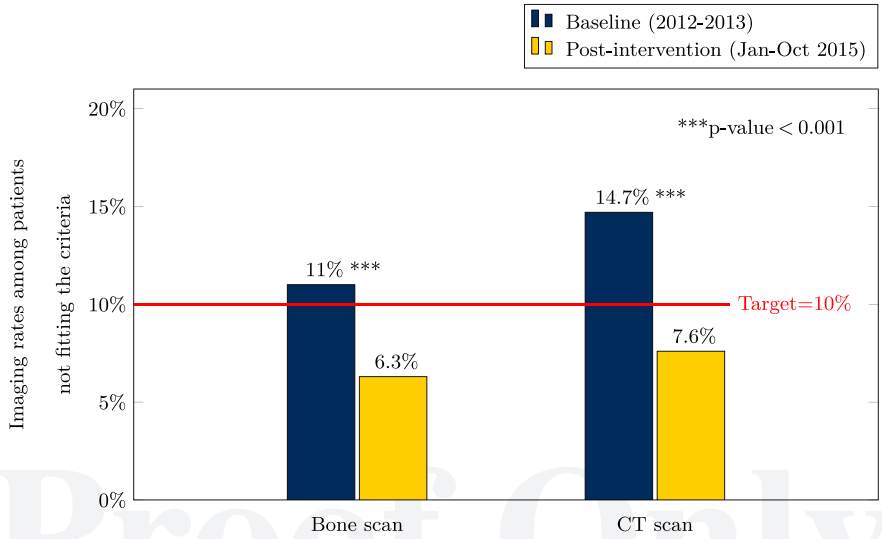
MUSIC members were provided with a toolkit including placards with the criteria (shown in Figure 8) and explanations for patients. After implementation, members also received comparative performance feedback that detailed how well their practice patterns correlated with the MUSIC Imaging Appropriateness Criteria.

After implementing this intervention in 2014, MUSIC measured postintervention outcomes from January to October 2015. The results showed an increase in the use of BSs and CT scans in patients meeting the criteria from 82% to 84% and from 74% to 77%, respectively. One of the underlying reasons why the imaging recommendations were not followed in some cases could be attributed to economic challenges faced by patients, such as high copays, depending on a patient's health coverage status; however, the full range of reasons is unknown. 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, MUSIC decreased the use of BSs 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, California (Hurley et al. 2016).

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 MUSIC 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 postimplementation results 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

Figure 9. (Color online) Avoidance of Low-Value Imaging Using MUSIC Criteria



***$p \leq 0.001$.

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's 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). Our publications were cited in the new National Comprehensive Cancer Network (2014; NCCN) guidelines and AUA guidelines (Sanda et al. 2017). 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 on the successes described above. For example, this work has led to a biopsy outcome prediction calculator, which has been implemented as a web-based decision support system called AskMUSIC (see https://askmusic.med.umich.edu/). Moreover, the approach we describe in this article has broad applicability to other disease for which imaging is used (e.g., other cancers like breast cancer, kidney cancer, thyroid cancer) and other gold-standard tests with binomial outcomes such as molecular biomarkers and biopsies.

Ideally, from a research perspective, the impact of the actual use of the intervention on health outcomes and cost-effectiveness should be studied through randomized clinical trials. However, in practice, such randomized trials are expensive and time-consuming. Moreover, there are important ethical issues associated with imaging patients without any clinical indication, such as exposure to harmful radiation. Imaging tests such as CT scans expose patients to significant amounts of radiation, although withholding imaging from patients in need could result in undetected metastatic cancer. Using observational clinical data, we provided a multistep framework for establishing evidence-based imaging recommendations and evaluated the potential impact of the recommendations using the predictive models that we developed and validated. A potential barrier to implementation in other regions is that the patient population seen in MUSIC might be different from patient populations in other practices; for example, the change in patient population might be the effect of a change in referral pattern by practitioners or genetic differences that impact the propensity for prostate cancer to metastasize. The validity of the predictive models we developed using MUSIC data need to be established in independent samples for a wider implementation, which is possible only in cases where sufficient data have been collected. Other states are starting to develop quality improvement collaboratives for prostate cancer, and we believe that our work provides a road map for broader implementation to inform health policy decisions and to improve patient care and population outcomes.

Acknowledgments

The authors thank Susan Linsell and the Michigan Urological Surgery Improvement Collaborative for their aid in this project. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Appendix A. Patient Characteristics

Variables	Bone scan: All patients without BS (n = 1,103)	Bone scan: All patients with BS (n = 416)	Bone scan: p-value	CT scan: All patients without CT scan (n = 1,737)	CT scan: All patients with CT scan (n = 643)	CT scan: p-value
Age (years)			0.02			0.17
Mean (median)	64.2 (64.4)	68.2 (67.7)		63.8 (64.0)	66.0 (66.0)	
Range	40.4–95.8	41.8–90.5		40.4–95.0	40.0–99.0	
Clinical stage, No. (\%)			<0.0001			<0.0001
T1	881 (79.9)	216 (51.9)		1,386 (79.8)	359 (55.8)	
T2	214 (19.4)	173 (41.6)		339 (19.5)	246 (38.3)	
T3/4	8 (0.7)	27 (6.5)		12 (0.7)	38 (5.9)	
PSA, ng/mL			0.003			<0.0001
Mean (median)	8.0 (5.2)	61.8 (7.7)		8.60 (5.2)	49.9 (7.7)	
Range	0.2–620.8	0.4–6,873.4		0.23–1,008.9	0.40–6,873.4	

Appendix A. (Continued)

Variables	Bone scan: All patients without BS (n = 1,103)	Bone scan: All patients with BS (n = 416)	Bone scan: p-value	CT scan: All patients without CT scan (n = 1,737)	CT scan: All patients with CT scan (n = 643)	CT scan: p-value
ln(PSA + 1)			<0.0001			<0.0001
Mean (median)	1.9 (1.8)	2.5 (2.2)		1.9 (1.8)	2.5 (2.2)	
Range	0.2–6.4	0.3–8.8		0.2–6.9	0.3–8.8	
Biopsy Gleason sum, No. (\%)			<0.0001			<0.0001
≤6	488 (44.2)	33 (7.9)		747 (43.0)	62 (9.6)	
3+4	439 (39.8)	105 (25.2)		671 (38.6)	174 (27.1)	
4+3	137 (12.4)	58 (13.9)		212 (12.2)	97 (15.1)	
8–10	39 (3.6)	220 (52.9)		107 6.2)	310 (48.2)	
Biopsy cores taken, No.			0.50			0.40
Mean (median)	12.5 (12.0)	12.9 (12.0)		12.5 (12.0)	12.7 (12.0)	
Range	4–82	1–78		2–82	1–78	
Positive cores, No.			0.0004			<0.0001
Mean (median)	3.2 (3.0)	6.3 (6.0)		3.3 (3.0)	6.2 (6.0)	
Range	0–20	1–16		1–20	1–16	
Positive cores, %			<0.0001			<0.0001
Mean (median)	26.4 (21.1)	51.2 (50.0)		27.0 (23.1)	50.4 (50)	
Range	0–100	3.1–100		2.44–100	3.13–100	

Appendix B. Cost-Sensitive Laplacian Kernel Logistic Regression

We begin by introducing the problem for KLR. 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 with linear models. To include nonlinear decision boundaries in our problem, we extend the construction from LR given in (2) to KLR by incorporating a nonlinear feature mapping into the decision function: $f(\mathbf{x}) = \Phi(\mathbf{x})\boldsymbol{\beta}$ (Zhu and Hastie 2005, Maalouf et al. 2011). Given a set of l imaged patients $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)\} \in (\mathbb{R}^n \times \{-1, 1\})^l$, the optimization problem for KLR becomes the following:

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

where $\boldsymbol{\beta} \in \mathcal{H}$ is the parameter we want to estimate. By (4) and the kernel trick, that is, $\langle \boldsymbol{\Phi}(\mathrm{x}), \boldsymbol{\Phi}(\mathrm{x}')\rangle = \mathbf{K}(\mathrm{x}, \mathrm{x}')$, the minimizer of (B.1) admits a representation of the form $\boldsymbol{\beta} = \sum_{i=1}^{l} \alpha_i \Phi(\mathbf{x}_i)$. Thus, we can write (B.1) as

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

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 ith element of the vector $\mathbf{K}\boldsymbol{\alpha}$.

We now extend the supervised KLR to the cost-sensitive semisupervised KLR via graph regularization as proposed by Belkin et al. (2006). Assume now that we are given a set of u unimaged patients $\{\mathbf{x}_j\}_{j=l+1}^{j=l+u}$ in addition to the labeled data. 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 $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j)\rangle$. Because 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 are 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 $d_i = \sum_{j=1}^{l+u} w_{ij}$. We define the weight matrix $\mathbf{W}$ by k nearest neighbors 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 define $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_{l+u})]^T$ and $\mathbf{L}$ as the Laplacian matrix of the graph given by $\mathbf{L} = \mathbf{D} - \mathbf{W}$. With the incorporation of the partial losses from (5), we consider the following cost-sensitive optimization problem:

$$\begin{aligned} f^* = \arg\min_{f \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} \Big[& \delta \mathbb{1}_{\{y_i = 1\}} \log\left(1 + e^{-f(\mathbf{x}_i)}\right) \\ & + (1 - \delta) \mathbb{1}_{\{y_i = -1\}} \log\left(1 + e^{f(\mathbf{x}_i)}\right) \Big] + \gamma_{\mathcal{H}} \|f\|_{\mathcal{H}}^2 \\ & + \gamma_{\mathcal{M}} \mathbf{f}^{\mathbf{T}} \mathbf{L} \mathbf{f}, \end{aligned} \tag{B.3}$$

where $\gamma_{\mathcal{H}}$ controls the complexity of f in space $\mathcal{H}$, and $\gamma_{\mathcal{M}}$ controls the complexity of the function in the intrinsic geometry of marginal distribution $\mathcal{P}_X$ (Belkin et al. 2006). We refer to the optimization problem in (B.3) as Cos-LapKLR. The extensions of standard regularization algorithms by solving the optimization problems (posed in (3)) for different choices of cost function L and regularization parameters $\gamma_{\mathcal{H}}$ and $\gamma_{\mathcal{M}}$ have been developed (Belkin et al. 2006).

As in (4), the representer theorem can be used to show that the solution to (B.3) 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 let $\mathbf{K}_L \in \mathbb{R}^{l \times l}$ be the kernel matrix for imaged patients. In order to express (B.3) 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, (B.3) can then be equivalently rewritten in a finite dimensional form as

$$\begin{aligned} H(\boldsymbol{\alpha}) = \min_{\boldsymbol{\alpha}\in\mathbb{R}^{l+u}} \frac{1}{2l}\Big[\delta \mathbf{1}(\mathbf{1}+\mathbf{y})^T \log\Big(\mathbf{1}+e^{-(\mathbf{K}_L\mathbf{P}_L\boldsymbol{\alpha})}\Big) + \\ + (1-\delta)\mathbf{1}(\mathbf{1}-\mathbf{y})^T \log\Big(\mathbf{1}+e^{(\mathbf{K}_L\mathbf{P}_L\boldsymbol{\alpha})}\Big)\Big] \\ + \gamma_{\mathcal{H}}\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} + \gamma_{\mathcal{M}}\boldsymbol{\alpha}^T\mathbf{K}\mathbf{L}\mathbf{K}\boldsymbol{\alpha}. \end{aligned} \tag{B.4}$$

The outline of the algorithm we propose for solving Cos-LapKLR is given in Algorithm B.1. It is natural to use the Newton–Raphson method to fit Cos-LapKLR because (B.4) 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 that uses a limited amount of computer memory (Byrd et al. 1995).

Algorithm B.1 (Cost-Sensitive Laplacian Kernel Logistic Regression)

Input: l labeled examples $\{(\mathrm{x}_i, y_i)\}_{i=1}^{l}$, u unlabeled examples $\{\mathrm{x}_j\}_{j=l+1}^{l+u}$

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 $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_{l+u})$ 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 α^* using (B.4) together with the LM-BFGS algorithm.

Step 6: Output function $f^*(\mathrm{x}) = \sum_{i=1}^{l+u} \alpha_i^* \mathbf{K}(\mathrm{x}_i, \mathrm{x})$.

Appendix C. 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 nonheuristic sampling methods are commonly used: ROS and RUS.

The synthetic minority oversampling technique (SMOTE) is a method of oversampling that 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 classes. Borderline-SMOTE was proposed to enhance the original concept by identifying the

Table C.1. Performance of RFs and AdaBoost for BS and CT Scan in 10 Independent Repetitions of Twofold CV

	BS (n = 416)				CT scan (n = 643)			
Models	Sensitivity	Specificity	AUC	*Brier*	Sensitivity	Specificity	AUC	*Brier*
RFs								
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
Borderline-SMOTE	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

Q: 45

Table C.1. (Continued)

	BS (n = 416)				CT scan (n = 643)			
Models	Sensitivity	Specificity	AUC	*Brier*	Sensitivity	Specificity	AUC	*Brier*
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
Borderline-SMOTE	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

Note. Sensitivity, specificity, and AUC are reported in percentages.

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 the 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, the condensed nearest neighbor rule (CNN; Hart 1968), and the neighborhood cleaning rule (NCR; Laurikkala 2001). In this work, we implement and test 10 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 twofold cross-validation on the development samples. The results from these experiments are summarized in Table C.1.

The experimental results indicate that the accuracy of classification rules on the BS and CT scan data sets developed by RFs and AdaBoost can be improved via model-independent data-driven approaches. For instance, the baseline RFs identifying patients with bone metastasis obtained a sensitivity of 24.97% and specificity of 98.05%, whereas RFs combined with RUS improved the sensitivity to 74.68% while reducing the specificity to 68.13%. RFs and AdaBoost combined with RUS achieved the highest sensitivity and AUC in both BS and CT scan data sets. These results clearly illustrate the inadequacy of the baseline RFs and AdaBoost in recognizing metastatic patients.

References

American Cancer Society (2019a) Key statistics for prostate cancer. https://www.cancer.org/cancer/prostate-cancer/about/key-statistics.html. Q: 46

American Cancer Society (2019b) Prostate cancer stages and other ways to assess risk. https://www.cancer.org/cancer/prostate-cancer/detection-diagnosis-staging/staging.html.

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.

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

Begg CB, Greenes RA (1983) Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* 39(1): 207–215.

Belkin M, Niyogi P, Sindhwani V (2006) Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Machine Learn. Res.* 7:2399–2434. Q: 47

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. *J. Clinical Epidemiology* 56(9):826–832.

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

Briganti A, Passoni N, Ferrari M, Capitanio U, Suardi N, Gallina A, Da Pozzo LF, 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. *Eur. Urology* 57(4):551–558.

Broder M, Gutierrez B, Cherepanov D, Linhares Y (2015) Burden of skeletal-related events in prostate cancer: Unmet need in pain improvement. *Support. Care Cancer* 23(1):237–247.

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

Cancer.Net (2019) Prostate cancer: Stages and grades. https://www.cancer.net/cancer-types/prostate-cancer/stages-and-grades. Q: 48

Carroll P, Albertsen PC, Greene K, Babaian RJ, Ballentine Carter H, Gann PH, Han M, et al. (2013) PSA testing for the pretreatment staging and posttreatment management of prostate cancer. https://www.auanet.org/guidelines/prostate-specific-antigen-(2009-amended-2013). Q: 49

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

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

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

Cox DR (1958) Two further applications of a model for binary regression. *Biometrika* 45(3/4):562–565.

Cruz JA, Wishart DS (2006) Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics* 2:59–77. Q: 50

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

Domingos P (1999) Metacost: A general method for making classifiers cost-sensitive. *Proc. Fifth ACM SIGKDD Internat. Conf. Knowledge*

Discovery Data Mining (Association for Computing Machinery, New York), 155–164.

Efron B, Tibshirani R (1997) Improvements on cross-validation: The 632+ bootstrap method. *J. Amer. Statist. Assoc.* 92(438):548–560.

Efron B, Tibshirani RJ (1994) *An Introduction to the Bootstrap* (CRC Press, Boca Raton, FL).

Elkan C (2001) The foundations of cost-sensitive learning. *Internat. Joint Conf. Artificial Intelligence*, vol. 17, 973–978.

Evgeniou T, Pontil M, Poggio T (2000) Regularization networks and support vector machines. *Adv. Comput. Math.* 13(1):1–50.

Friedman J, Hastie T, Tibshirani R (2000) Additive logistic regression: A statistical view of boosting. *Ann. Statist.* 28(2):337–407.

Friedman J, Hastie T, Tibshirani R (2001) *The Elements of Statistical Learning*, vol. 1 (Springer, Berlin).

Fuchsjäger M, Shukla-Dave A, Akin O, Barentsz J, Hricak H (2008) Prostate cancer imaging. *Acta Radiologica* 49(1):107–120.

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. Huang DS, Zhang XP, Huang GB, eds. *Internat. Conf. Intelligent Comput.* (Springer, Berlin), 878–887.

Harrell F, Lee KL, Mark DB (1996) Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statist. Medicine* 15(4):361–387.

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

He H, Garcia EA (2009) Learning from imbalanced data. *IEEE Trans. Knowledge Data Engrg.* 21(9):1263–1284.

Heidenreich A, Bastian PJ, Bellmunt J, Bolla M, Joniau S, van der Kwast T, Mason M, et al (2014) EAU guidelines on prostate cancer. Part II: Treatment of advanced, relapsing, and castration-resistant prostate cancer. *Eur. Urology* 65(2):467–479.

Hricak H, Choyke PL, Eberhardt SC, Leibel SA, Scardino PT (2007) Imaging prostate cancer: A multidisciplinary perspective. *Radiology* 243(1):28–53.

Hsu CW, Chang CC, Lin CJ (2003) A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, Taipei, Taiwan.

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

Kim L, Min M, Roos D, Nguyen L, Yeoh E (2015) Are staging investigations being overused in patients with low and intermediate risk prostate cancer? *J. Medical Imaging Radiation Oncology* 59(1):77–81.

Kimeldorf G, Wahba G (1971) Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* 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. *Comput. Structural Biotechn. J.* 13:8–17.

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

Lavery HJ, Brajtbord JS, Levinson AW, Nabizada-Pace F, Pollard ME, Samadi DB (2011) Unnecessary imaging for the staging of low-risk prostate cancer is common. *Urology* 77(2):274–278.

Lemaître G, Nogueira F, Aridas CK (2017) Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Machine Learn. Res.* 18(17):1–5.

Li Y, Kwok JTY, Zhou ZH (2010) Cost-sensitive semi-supervised support vector machine. *Proc. Natl. Conf. Artificial Intelligence*, vol. 1, 500.

Lin EC (2010) Radiation risk from medical imaging. *Mayo Clinic Proceeding* 85(12):1142–1146.

Little RJ (1988) A test of missing completely at random for multivariate data with missing values. *J. Amer. Statist. Assoc.* 83(404): 1198–1202.

Liu A, Jun G, Ghosh J (2009) Spatially cost-sensitive active learning. *Proc. SIAM Internat. Conf. Data Mining* (Society for Industrial and Applied Mathematics, Philadelphia), 814–825.

Liu XY, Zhou ZH (2006) The influence of class imbalance on cost-sensitive learning: An empirical study. *6th Internat. Conf. Data Mining* (Institute of Electrical and Electronics Engineers, Piscataway, NJ), 970–974.

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

Maloof MA (2003) Learning when data sets are imbalanced and when costs are unequal and unknown. *Internat. Conf. Machine Learn Workshop Learn. Imbalanced Datasets* II, vol. 2.

Margineantu DD (2005) Active cost-sensitive learning. *Internat. Joint Conf. Artificial Intelligence*, vol. 5, 1622–1623.

Masnadi-Shirazi H, Vasconcelos N (2010) Risk minimization, probability elicitation, and cost-sensitive SVMs. *Proc. 27th Internat. Conf. Machine Learn.*, 759–766.

McCarthy K, Zabar B, Weiss G (2005) Does cost-sensitive learning beat sampling for classifying rare classes? *Proc. 1st Internat. Workshop Utility-Based Data Mining* (Association for Computing Machinery, New York), 69–77.

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. *Med. Decision Making* 13(1):49–57.

Mottet N, Bellmunt J, Briers E, Association EU, et al. (2014) Guidelines on prostate cancer. *Eur. Urology* 65(1):124–137.

National Comprehensive Cancer Network (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, Oxford, UK).

Pockett R, Castellano D, McEwan P, Oglesby A, Barber B, Chung K (2010) 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. *Eur. J. Cancer Care (Engl.)* 19(6):755–760.

Prasad KN, Cole WC, Hasse GM (2004) Health risks of low dose ionizing radiation in humans: A review. *Experiment. Biol. Medicine* 229(5):378–382.

Prasad SM, Gu X, Lipsitz SR, Nguyen PL, Hu JC (2012) Inappropriate utilization of radiographic imaging in men with newly diagnosed prostate cancer in the united states. *Cancer* 118(5):1260–1267.

Qi Z, Tian Y, Shi Y, Yu X (2013) Cost-sensitive support vector machine for semi-supervised learning. *Procedia Comput. Sci.* 18:1684–1689.

Qin Z, Zhang S, Liu L, Wang T (2008) Cost-sensitive semi-supervised classification using CS-EM. *8th IEEE Internat. Conf. Comput. Inform. Tech.* (Institute of Electrical and Electronics Engineers, Piscataway, NJ), 131–136.

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.

Sanda MG, Chen RC, Crispino T, Freedland S, Nelson M, Reston J, Rodrigues G, Sandler HM, Taplin ME, Cadeddu JA (2017)

Clinically localized prostate cancer: AUA/ASTRO/SUO guideline. *Risk* 6:27.

Schapire RE (2003) The boosting approach to machine learning: An overview. Denison DD, Hansen MH, Holmes CC, Mallick B, Yu B, eds. *Nonlinear Estimation and Classification* (Springer, New York), 149–171.

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

Smith-Bindman R, Lipson J, Marcus R, Kim KP, Mahesh M, Gould R, De González AB, Miglioretti DL (2009) Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. *Arch. Internal Medicine* 169(22):2078–2086.

Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG (2003) Internal and external validation of predictive models: A simulation study of bias and precision in small samples. *J. Clinical Epidemiology* 56(5):441–447.

Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans M, Vergouwe Y, Habbema JDF (2001) Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *J. Clinical Epidemiology* 54(8):774–781.

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, et al (2007) Guideline for the management of clinically localized prostate cancer: 2007 update. *J. Urology* 177(6):2106–2131.

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

Tokan F, Türker N, Yıldırım T (2006) ROC analysis as a useful tool for performance evaluation of artificial neural networks. Kollias S, Stafylopatis A, Duch W, Oja E, eds. *Artificial Neural Networks.* Lecture Notes in Computer Science, vol. 4132 (Springer, Berlin), 923–931.

Tombal B, Lecouvet F (2012) Modern detection of prostate cancer's bone metastasis: Is the bone scan era over? *Adv. Urology* 2012:893193.

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

Vapnik V (2013) *The Nature of Statistical Learning Theory* (Springer, New York).

Veropoulos K, Campbell C, Cristianini N (1999) Controlling the sensitivity of support vector machines. *Proc. Internat. Joint Conf. Artificial Intelligence*, 55–60.

Von Luxburg U (2007) A tutorial on spectral clustering. *Statist. Comput.* 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 Trans. Systems Man Cybernetics* 2(3): 408–421.

Yong C, Onukwugha E, Mullins CD (2014) Clinical and economic burden of bone metastasis and skeletal-related events in prostate cancer. *Current Opinion Oncology.* 26(3):274–283.

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

Zhou XH, McClish DK, Obuchowski NA (2009) *Statistical Methods in Diagnostic Medicine*, vol. 569. (John Wiley & Sons, Hoboken, NJ).

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

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

Zhu J, Hastie T (2005) Kernel logistic regression and the import vector machine. *J. Comput. Graphical Statist.* 14(1):185–205.

Zhu X (2007) Semi-supervised learning literature survey. Technical report, Computer Science, University of Wisconsin–Madison, Madison.

Zhu X, Goldberg AB (2009) *Introduction to Semi-Supervised Learning: Synthesis Lectures on Artificial Intelligence and Machine Learning* (Morgan & Claypool, Williston, VT).

NUMBER 1 OF 6

AUTHOR QUERIES

DATE 9/24/2020

JOB NAME OPRE

ARTICLE 20202020

QUERIES FOR AUTHOR Merdan et al.

THIS QUERY FORM MUST BE RETURNED WITH ALL PROOFS FOR CORRECTIONS

Q: 1_Please confirm the funding information and verify that all funding is listed.

Q: 2_Your color figures appear in color in your page proof to simulate their presentation in color online, but they will be converted to grayscale in the print version of your article because you have not requested color processing for print. Please (a) check your figures to confirm that they will be sufficiently clear in grayscale presentation in the print version of your article and (b) modify your figure legends as necessary to remove any references to specific colors in the figure.

Q: 3_Please confirm that subject classifications are correct as set.

Q: 4_Please provide a short blurb of 1-2 sentences summarizing your article for the table of contents.

Q: 5_Please confirm that the article title, author names, affiliations, and email addresses are set correctly. If applicable, please provide author ORCID numbers.

Q: 6_Please confirm that keywords are correct as set.

Q: 7_Please confirm that heading levels are correct as set.

Q: 8_Please verify that all displayed equations and in-text math notations are set correctly.

Q: 9_Please confirm/correct the edit to "Therefore, accurate staging is crucial for the clinical management of prostate cancer, from possible cure. . .."

Q: 10_Per INFORMS style, stacked equations within paragraph text should be converted to inline equations if possible. Please "slash down" any stacked equations that appear within paragraph text, modifying them as necessary to be represented accurately as inline equations.

Q: 11_Per INFORMS style, all variables should be italic and all vectors should be bold. Please confirm that all terms have been formatted properly throughout.

Q: 12_Please confirm the edit to "Conventional imaging tests for prostate cancer staging include bone scans (BSs). . .."

Q: 13_Please confirm the sentence "In such cases, patients are subjected to treatments, such as radical prostatectomy (surgical removal of the prostate). . .."

Q: 14_Please confirm the URL "http://musicurology.com/."

Q: 15_Please review the sentence "This section describes the development and testing of predictive models. . .." Would "the probability of an imaging test being positive" be correct?

Q: 16_Please confirm/correct the edit to "and the maximum percentage of core involvement."

Q: 17_Please spell out "TNM" at only use.

Q: 18_Please confirm "biopsy GS (≤3 + 4, 4 + 3, or 8 − 10." Is "8–10" meant?

Q: 19_Please confirm the edit to "We used a logarithmic transformation of PSA scores because the distribution of PSA scores was highly skewed." Note that "since" has been changed to "because" throughout, where applicable, per journal style.

Q: 20_Equations in the text and appendices have been renumbered according to journal style. Please confirm all edits.

Q: 21_Please confirm that Algorithms 1 and B.1 (renumbered as such per journal style) are set, aligned, and punctuated correctly.

Q: 22_Please confirm/correct the edit to "The change in the AUC between the internal and external validations for BS and CT models was not significant."

Q: 23_Please confirm edits to the sentence "Given a set of labeled data, $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)\} \in (\mathbb{R}^n \times \{-1, 1\})^l$. . .."

Q: 24_Please confirm the edit to "The same algorithmic framework is utilized in many regression and classification schemes, such as support vector machines. . .." Would "regularized least squares regression" be correct?

Q: 25_Journal style is to use plural abbreviations for plural terms. Please confirm all edits to "RFs" throughout.

Q: 26_Please define "AdaBoost" at first use if possible.

Q: 27_Please confirm/correct the edits to "the cost of misclassifying the negative class into the positive class and the cost of misclassifying the positive class . . ." and similar edits in the subsequent sentence.

Q: 28_Please confirm/correct "minimum node size (5)." Is "minimum number of nodes" meant?

Q: 29_Please confirm the sentence "Our initial experiments explored how the cost ratio, δ. . .."

Q: 30_The term *verification bias* has been set in italics at the first use in text only, per journal style.

Q: 31_Please confirm the edit to "To obtain unbiased estimates . . . Begg and Greenes (1983) developed. . .." The abbreviation "B&G," used only once, has been removed.

Q: 32_Please confirm/correct all edits to "This method uses the observed proportions of patients with and without the disease among the verified patients. . .. The proportions are then combined to obtain a complete two-by-two table. . .."

Q: 33_Please confirm/correct the edits to the sentence “We further obtain unbiased estimates of $\mathbb{P}(G+)$ and $\mathbb{P}(G-)$….”

Q: 34_“The MUSIC collaborative” has been changed to “MUSIC” throughout, as “collaborative” is part of its definition. Please confirm.

Q: 35_Please confirm/correct the edit to “To define the criteria to be considered in the objective function, let p_i …, and let it be estimated from an LR model.”

Q: 36_Please confirm the link “https://youtu.be/FEIxb_HRHAA.”

Q: 37_Journal style is to avoid replacing words with math symbols in running text where possible. Can “greater than” and “less than” be written out in words in “Although these values are not >95%” and “Both of these values are below their goal of performing imaging in <10\%” below?

Q: 38_Please confirm the URL “https://askmusic.med.umich.edu/.”

Q: 39_Please confirm/correct the sentence “Imaging tests such as CT scans expose patients to significant amounts of radiation.” Journal style is to use “while” only in a time context.

Q: 40_Please confirm the acknowledgments.

Q: 41_Please indicate where italics are needed for variables in Appendix A. Alternatively, would "Characteristic" be an appropriate replacement for heading "Variables"?

Q: 42_In Appendix A, should “Gleason sum” be set as “Gleason score”?

Q: 43_In the sentence “By (4) and the kernel trick ...” please reword “$\boldsymbol{\beta} = \sum_{i=1}^{l} \alpha_i \Phi(\mathbf{x}_i)$” to set the limits to the right of the summation sign, for space considerations, or mark to set this term as a display.

Q: 44_Please confirm “We define the weight matrix $\mathbf{W}$ by k nearest neighbors.”

Q: 45_Please confirm/correct edits to "Borderline-SMOTE" in Table C.1, per the form given in the text.

Q: 46_Confirm the URL and provide the date accessed for American Cancer Society 2019a and b.

Q: 47_Provide the issue number for Belkin et al. 2006 if possible.

Q: 48_Please confirm the URL for Cancer.Net and provide the date accessed.

Q: 49_Confirm the author line and URL for Carroll et al. 2013 and provide the date accessed.

Q: 50_Please provide the issue number for Cruz and Wishart 2006 and Kourou et al. 2015 if possible.

Q: 51_Please provide any applicable editor names for Domingos 1999 and Laurikkala 2001.

Q: 52_Provide the publisher name and location and any applicable editor names for Elkan 2001.

Q: 53_Please confirm the author line for Friedman et al. 2000.

Q: 54_Confirm/correct the publisher location for Friedman et al. 2001.

Q: 55_Please confirm/correct all publication details for Han et al. 2005.

Q: 56_Please confirm the title and issue number for Harrell et al. 1996.

Q: 57_Please provide the publisher and its location, any applicable editor names, and complete page range for Li et al. 2010.

Q: 58_For Liu and Zhou 2006, Liu et al. 2009, and McCarthy et al. 2005, please confirm the publisher and its location and provide any applicable editor names.

Q: 59_For Maloof 2003, Margineantu 2005, and Masnadi-Shirazi and Vasconcelos 2010, please confirm the name of the conference and provide the publisher name and location and any applicable editor names. For Maloof 2003, also provide the complete page range.

Q: 60_Please correct the title, author line, and/or publication details for Mottet et al. 2014. The given publication details are for Heidenreich et al., "EAU Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent—Update 2013." Please also provide at least the first 11 author names, and clarify "Association EU" if retained.

Q: 61_Please confirm the accuracy of the title of the page cited for National Comprehensive Cancer Network (2014), and confirm/correct the URL and provide the date accessed.

Q: 62_For Qin et al. 2008, please confirm the publisher and its location and provide any applicable editor names.

Q: 63_Please confirm/correct the title and publication details for Sanda et al. 2017. Provide the volume, issue, and complete page range.

Q: 64_Please confirm/correct all publication details for Schapire 2003.

Q: 65_For Sindhwani et al. 2005, please provide the publisher and its location and any editor names and page numbers.

Q: 66_Please confirm all publication details added for Tokan et al. 2006.

Q: 67_Please confirm/correct the volume and article number for Tombal and Lecouvet 2012.

Q: 68_For Veropoulos et al. 1999, please provide the publisher name and location and any applicable editor names.

Q: 69_Confirm the page range added for Zhu and Hastie 2005.

Q: 70_Please confirm/correct the book title and publisher/location for Zhu and Goldberg 2009.

Q: 71_Please check that all tables and figures, including their titles and notes (which may contain edits) are set correctly.

Q: 72_Please describe the error bars in Figure 2 in a figure note.

Q: 73_Please confirm/correct the notes added below Figures 5 and 6.

Q: 74_Please confirm/correct the edits to the Table 1 notes.

Q: 75_Please confirm the citation added in Table 4.

Q: 76_Please edit each biography so that it includes only current position, research interests, and awards. Each biography should be <500 characters including spaces.

Selin Merdan received the M.S. degree in 2012 and the Ph.D. degree in 2018 from the University of Michigan, Ann Arbor, all in industrial and operations engineering. She was a postdoctoral scholar at Georgia Institute of Technology, Atlanta, 2018 and 219. She is currently an Associate Director of Machine Learning and Data Science at Value Analytics Labs. Her research interests include using optimization, machine learning and health economics to develop and translate innovative ideas into evidence-informed practices for enhancing the quality and value of patient care. She is a recipient of the Rackham Predoctoral Fellowship and the Richard and Eleanor Towner Prize for Outstanding PhD Research from the University of Michigan, the INFORMS Doing Good with Good OR Award, and the honorable mention for the INFORMS George B. Dantzig Dissertation Award.

Brian Denton is Chair of the Department of Industrial and Operations Engineering at the University of Michigan. His research interests are in data-driven sequential decision making and optimization under uncertainty with applications to medicine. Before joining the University of Michigan he worked at IBM, Mayo Clinic, and North Carolina State University. His honors and awards include being an INFORMS Fellow, and winning the National Science Foundation Career Award, the INFORMS Service Section Prize, the INFORMS Daniel H. Wagner Prize, the Institute of Industrial Engineers Outstanding Publication Award, and the Canadian Operations Research Society Best Paper Award. He served on the Editorial Boards of Health Systems, IIE Transactions, Interfaces, Manufacturing & Service Operations Management, Medical Decision Making, Operations Research, Optimization in Engineering, and Production and Operations Management. He is past Chair of the INFORMS Health Applications Section, and he served as Secretary of INFORMS, and President of INFORMS.

Christine Barnett, PhD, is a Senior Research Health Economist at RTI Health Solutions. Her research interests are in developing decision analytic models such as cost-effectiveness, cost-utility, and budget-impact models to evaluate the health and economic implications of health technologies, including pharmaceutical and medical device products. Dr. Barnett has developed models and analyses in the areas of oncology, degenerative disc disease, thrombocytopenia, depression, and rare pediatric diseases. Prior to joining RTI Health Solutions, Dr. Barnett worked at the Mayo Clinic studying patient adherence to multiple medications in the context of type 2 diabetes. In addition, she worked on a grant for the Centers for Disease Control and Prevention, where she developed models and software used to determine the commercial distribution system for dispensing antivirals during an influenza pandemic. She is a recipient of the INFORMS Doing Good with Good OR Award and the National Science Foundation Graduate Research Fellowship. Her research has been presented at numerous professional conferences and published in peer-reviewed journals, including *Medical Decision Making*,

Cancer, *Urology*, *BJU International*, and *Journal of Medical Economics*. James Montie, MD, is Professor Emeritus in the Department of Urology at the University of Michigan and was Chairman of the Department from 1997-2007. His urology residency was at the Cleveland Clinic and oncology fellowship at Memorial Sloan-Kettering Cancer Center. He has been on the faculty at the Cleveland Clinic and Wayne State University prior to joining the University of Michigan in 1995. Dr. Montie initiated the Division of Health Services Research in the Department of Urology and is the founding Co-Director of the Michigan Urological Surgery Improvement Collaborative (MUSIC), a quality improvement collaborative that now includes 90% of the urologists in the state of Michigan. David Miller currently serves as Chief Clinical Officer for the University of Michigan Hospital and Frankel Cardiovascular Center. Working in partnership with the UH/CVC Executive Director and Chief Nursing Officer, Dr. Miller provides oversight and direction to ensure quality and safety in clinical programs and patient care, improve patient experience, enhance staff engagement, and optimize capacity management within the two hospitals. Dr. Miller is a Professor in the Department of Urology and maintains an active role in providing direct patient care. His clinical practice focuses on the diagnosis and management of patients with prostate and kidney cancer.

In addition to his clinical practice, Dr. Miller serves as Director of the Michigan Urological Surgery Improvement Collaborative (MUSIC). Funded by Blue Cross Blue Shield of Michigan (BCBSM), MUSIC is a consortium of more than 40 urology practices aiming to improve the quality and cost-efficiency of prostate cancer care in the state of Michigan. Dr. Miller also has a broad background in health services research, including substantial experience using claims data and formal training in the advanced statistical methods used in observational data analyses. With longitudinal funding from the Agency for Healthcare Research & Quality and the National Cancer Institute, Dr. Miller's empirical research agenda focuses on comparative effectiveness research, physician-led collaborative quality improvement, and understanding the relationship between physician organizations, integrated delivery systems, and the quality and cost of specialty care.