Statistics
in Medicine WILEY

Uncertainty in lung cancer stage for survival estimation via set-valued classification

Savannah Bergquist[1] | **Gabriel A. Brooks[2]** | **Mary Beth Landrum[3]** | **Nancy L. Keating[3]** | **Sherri Rose[4]**

[1]Haas School of Business, University of California, Berkeley, Berkeley, California, USA

[2]The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine, Hanover, New Hampshire, USA

[3]Department of Health Care Policy, Harvard Medical School, Boston, Massachusetts, USA

[4]Department of Health Policy and Center for Health Policy, Stanford University, Stanford, California, USA

Correspondence
Savannah Bergquist, Haas School of Business, University of California, Berkeley, 2220 Piedmont Ave. Office F461, Berkeley, CA 94720, USA.
Email: bergquist@berkeley.edu

The difficulty in identifying cancer stage in health care claims data has limited oncology quality of care and health outcomes research. We fit prediction algorithms for classifying lung cancer stage into three classes (stages I/II, stage III, and stage IV) using claims data, and then demonstrate a method for incorporating the classification uncertainty in survival estimation. Leveraging set-valued classification and split conformal inference, we show how a fixed algorithm developed in one cohort of data may be deployed in another, while rigorously accounting for uncertainty from the initial classification step. We demonstrate this process using SEER cancer registry data linked with Medicare claims data.

KEYWORDS
cancer, classification, conformal inference, survival analysis

1 | INTRODUCTION

Understanding generalizability across populations is important for answering applied health questions,[1] especially in the context of prediction algorithms.[2] In many cases, prediction tools are developed for a specific cohort or population but have limited practical applicability because it is difficult to rigorously assess the uncertainty

that comes from deploying a fixed prediction algorithm across settings. Population-level oncology research in the United States is an example of this phenomenon. There is great interest in predicting cancer stage using health care claims data because stage—an indicator of disease severity not included in claims data—is a key variable in understanding whether patients received appropriate treatments and for assessing health outcomes, including survival.[3,4] The availability of valid algorithms to identify cancer stage using only claims data would allow for timely assessments of quality and outcomes as well as comparative effectiveness research in non-clinical trial settings.

However, in applied classification settings, it can be a particular challenge to convey uncertainty. Often, a classification or risk prediction tool is developed with a focus on in-sample and validation sample performance measures, but little or no consideration is given to the use of the predicted values in subsequent estimation tasks.[5-9] Predicted values are commonly treated as observed data in downstream analyses. Depending on the specific applied context, this can lead to bias, erroneously deflated SEs, and decreased interpretability.[10,11]

To the best of our knowledge, there is no method that incorporates cancer stage classification uncertainty in survival estimation. The goal of this study is to develop a method that bridges between presenting raw probabilities—which may reflect our best sense of epistemic uncertainty albeit difficult to act upon in practice—and treating predicted class labels as ground truths. Using SEER registry data linked to Medicare claims data, we develop prediction algorithms to classify patients as having stage I/II, stage III, or stage IV lung cancer. After evaluating predictive performance, we then examine survival outcomes stratified by these stage classes. Motivated by this real-world applied problem of identifying stage for use in survival estimation, we propose an approach that leverages set-valued classification, split conformal inference, and resampling to produce sets of labels that simplify underlying classification uncertainty and translate it into outcomes estimation. We describe this procedure and compare it to a naive classification and outcomes estimation method without consideration of prediction uncertainty, and a naive method with uncertainty incorporated via resampling. In a simulation study, we explore conditions under which the standard practice approach differs from the bootstrap-based alternatives, and show how uncertainty may vary across class labels. Both bootstrap methods are an improvement over standard practice in terms of taking steps to more comprehensively communicate uncertainty in applied settings. Our applied data example shows that in practice the simpler, naive bootstrap tool may perform as well as the more complex weighted labeling bootstrap procedure.

1.1 | Related work on prediction and uncertainty

Multiple imputation and the measurement error modeling literatures are frequently applied in settings where a key variable of interest is missing for some or all subjects. Multiple imputation commonly seeks to generate values when a variable is missing for a subset of subjects in a data set. Covariate error measurement models often rely on some portion of observed data or external information pertaining to the variables measured with error and require assumptions about the nature of the error (eg, independent and additive). In cancer stage classification with claims data—and many other settings across medicine and health services research—the variable to be predicted is systematically missing for the entire analysis sample. In our proposed approach we avoid the assumption that errors are independent of the predicted values, and we do not require downstream users to observe any values for the variable to be predicted.

Wang et al[10] propose a method for correcting estimates, SEs, and test statistics in settings where the research question of interest focuses on the relationship between a covariate and outcome, but the outcome must first be predicted. They term inference with predicted outcomes "post-prediction inference." Their correction is based on the relationship between the observed and predicted outcome in a test data set. This post-prediction inference method is flexible but it is not developed for cases where the predicted variable is not the subsequent target outcome.

In clinical risk prediction, a popular method for communicating uncertainty is generating an uncertainty score to accompany the predicted values.[12,13] The aim of this work is often to provide a measure of uncertainty to facilitate clinical decision-making, rather than using the predicted values and associated uncertainty measure in statistical analysis. However, these methods share a high-level goal with our proposed method and post-prediction inference: improve the effectiveness of prediction tools for applied problem solving.

1.2 | Conformal inference and related methods

The conformal inference framework was developed by Vovk et al[14] to provide an unbiased solution for obtaining a prediction interval around a new observation drawn from the same distribution as a set of training data. Note that a naive solution to creating a prediction interval based on ranking the residuals from the training data is in practice unlikely to obtain target coverage, meaning the probability that a new test point lies in the prediction interval may be less than $1-\alpha$, where α is the specified error level. Full conformal inference is computationally expensive because it requires refitting the prediction algorithm, and recalculating and reordering the residuals for every new data point.

Split conformal inference separates the fitting and ranking steps so that the fitting only occurs as frequently as the data are split (ie, with a single split, the prediction algorithm is fit once and the residuals are ranked once). Papodopoulos,[15] Vovk,[16] and Lei et al[17] show that if the model is correctly specified, split conformal inference provides near-optimal conditional coverage; in cases where it is not, meaningful, marginal coverage may still be obtained. Health care data sets are typically subject to some gradual shifts in treatment practices and population demographics,[18] and these types of drifts may or may not threaten assumptions about the data distribution over limited time periods. However, investigator knowledge is critical for understanding events that can cause dramatic shifts and violate exchangeability (eg, changes in medical coding regulations or the introduction of a blockbuster drug). Although split conformal inference can lead to wider prediction intervals than methods that use a greater number of splits, it is less computationally expensive and has been developed to work with classification outcomes.[19,20] In this article, we deploy split conformal inference to leverage these characteristics and extend it for purposes of generalizability.

Although they are not directly applicable to our real-world data application of cancer stage classification and survival estimation because they do not produce label sets, we briefly describe two further, related approaches in the literature that produce prediction intervals for regression settings. Cross conformal inference[21,22] is a natural extension of split conformal inference: instead of dividing the data in two halves, the data are divided into V folds of equal size. Cross conformal inference has no coverage guarantees when V is large. Jackknife+ is a recently introduced method that uses leave-one-out predictions for the prediction interval.[23] Barber et al[23] also propose a method called CV+, which yields larger intervals than jackknife+ but requires less computation. We did not pursue adapting these approaches in this study because we are focused on an applied classification setting.

2 | METHODS

We will compare three methods, which are summarized conceptually in Figure 1. In the naive standard practice approach, the prediction algorithm is fit in the development data and applied to a validation data set, yielding a single label per observation that is used in subsequent outcomes analysis. In our data setting, the naive approach is to apply a fitted classification algorithm in the validation data, assign lung cancer stage based on the greatest predicted class probability, and then use this stage label to estimate survival. Under this naive approach, no special consideration is given to the uncertainty from predicting class labels. In the naive bootstrap method, uncertainty is incorporated by bootstrapping the validation data. This resampling allows for bootstrap-based confidence intervals around classification performance measures and outcomes estimation. Finally, the weighted labeling bootstrap procedure divides the development data into two halves to leverage split conformal inference (discussed in more detail below). The fitted algorithm and label thresholds are then applied to the validation data to assign a set of plausible labels for each individual observation. The validation data are resampled and a single label is selected from each label set and used in outcomes estimation. Thus the weighted labeling approach captures classification and outcomes estimation uncertainty via set-valued classification (the label sets) and resampling.

2.1 | LABEL classifier and weighted bootstrap

Our data structure is defined as $O = (X, Y)$, where X is a feature vector and $Y = \{1, \ldots, K\}$ is the mutually exclusive label space. In the multiclass setting, we seek to apply a label $y \in Y$ to each observation based on $x \in X$. We set $K = 3$; in our applied data analysis, lung cancer stage I/II, stage III, and stage IV map to labels 1-3, respectively. Although each observation has a single true label, our goal is to assign a *set* of plausible labels in order to provide more information about the uncertainty of the classification process. To do so, we employ a set-valued classifier proposed by Sadinle et al,[20] the
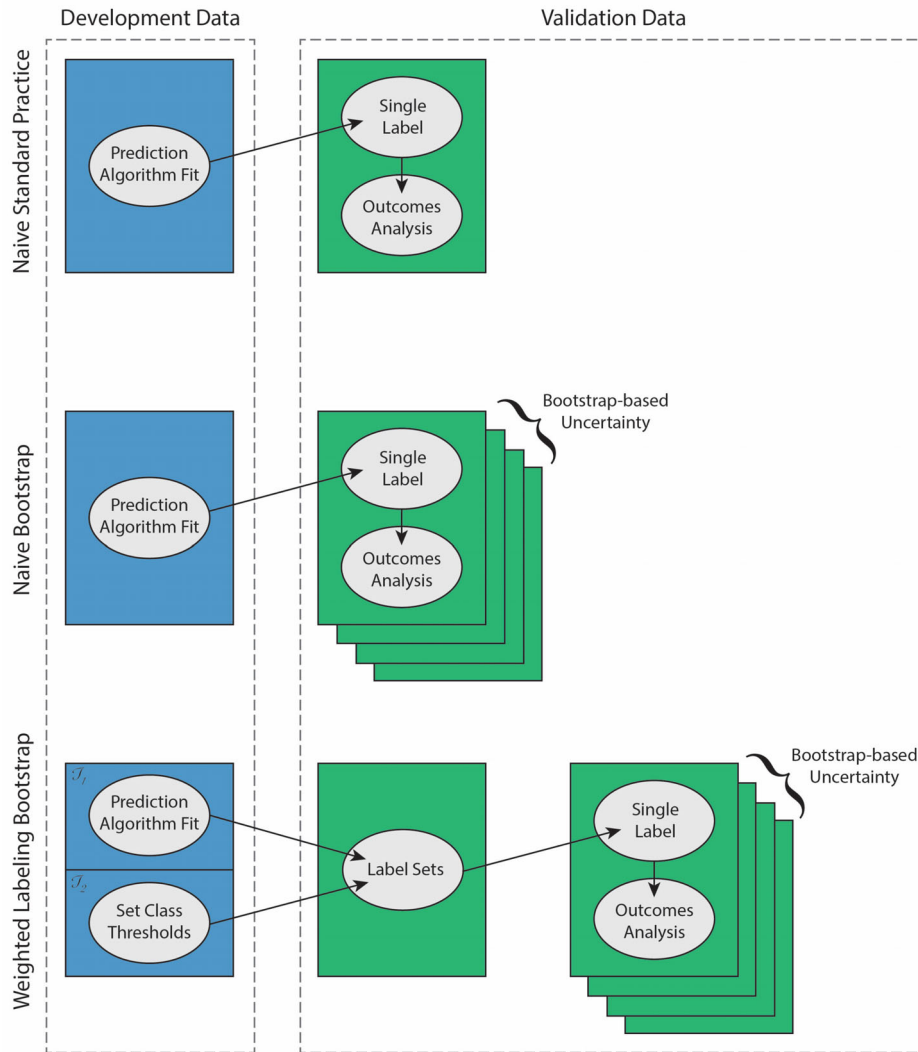
FIGURE 1 Conceptual overview of methods

least ambiguous with bounded error levels (LABEL) classifier. In combination with split conformal inference, the LABEL classifier yields distribution-free, finite sample coverage levels with minimal label ambiguity when the assumption of exchangeability holds. To estimate each y, we require an initial estimate of the conditional probability function, $P(y|x)$, and class-specific thresholds for converting probabilities to labels, $\{t_y\}_{y=1}^K$. Any conventional estimator of $P(y|x)$ can be plugged in to the LABEL classifier; in our simulation study we consider a multinomial logistic regression and in our data analysis we include a range of estimators.

We desire equal coverage across all three lung cancer stage classes, and therefore pre-specify individual class error levels of $\{\alpha_y\}_{y=1}^K = 0.10$. As described in Sadinle et al,[20] in the optimal procedure, the set-valued classifier, $\mathbf{H}(X)$, is a subset of the label space and can be represented by a collection of sets mapping the feature inputs to labels: $C_y = \{x \in X : y \in \mathbf{H}(X)\}$. The estimated sets for each class label, $\widehat{C}_y = \{x : \widehat{P}(y|x) \geq \widehat{t}_y\}$, are based on the estimated conditional probabilities, $\widehat{P}(y|x)$, and estimated thresholds, $\widehat{t}_y$.

The coverage for class y is based on the collection of estimated sets, $\widehat{C}_y$, that contain the true class label:

$$\widehat{\text{coverage}}_y(\widehat{t}_y) = \frac{\sum_{i=1}^n I(X_i \in \widehat{C}_y)I(Y_i = y)}{\sum_{i=1}^n I(Y_i = y)},$$

where i indexes n independent observations from O. The threshold $\widehat{t}_y$ is estimated:

$$\widehat{t}_y = \max_{i:Y_i=y} [\widehat{P}(Y_i|X_i) : \widehat{\text{coverage}}\{\widehat{P}(Y_i|X_i)\} \geq 1 - \alpha_y].$$

In an applied, finite sample setting with split conformal inference, the data are split in two halves, indexed by $\mathcal{I}_1$ and $\mathcal{I}_2$. For our application, the data are first divided into a development and validation cohort, and the development cohort is split into $\mathcal{I}_1$ and $\mathcal{I}_2$. The $\mathcal{I}_1$ subset is used to estimate $\widehat{P}(y|x)$. We partition $\mathcal{I}_2$ into K groups corresponding to each label class: $\mathcal{I}_{2,y} = \{i \in \mathcal{I}_2 : Y_i = y\}$. Within the given partition $\mathcal{I}_{2,y}$, thresholds $\widehat{t}_y$ are estimated according to pre-specified error levels α_y:

$$\widehat{t}_y = \min_{i \in \mathcal{I}_{2,y}}\left\{ \widehat{P}(Y_i|X_i) : \sum_{j \in \mathcal{I}_{2,y}} I\{\widehat{P}(Y_j|X_j) \leq \widehat{P}(Y_i|X_i)\} > (|\mathcal{I}_{2,y}| + 1)\alpha_y - 1 \right\},$$

where j indexes over $\mathcal{I}_2$ for comparison of each predicted probability to the ranked predicted probabilities. We then obtain the sets $\widehat{C}_y = \{x : \widehat{P}(y|x) \geq \widehat{t}_y\}$ so that $\{\widehat{C}_y\}_{y=1}^K$ is the split conformal set-valued classifier based on the plug-in estimator $\widehat{P}(y|x)$. Instances where more than one label is assigned to an observation are called ambiguous sets, and those with no assigned labels are null sets. With α_y set equally across classes, a higher threshold for a given class y indicates a greater number of higher predicted probabilities generated by $\widehat{P}(y|x)$. We apply the classifier—based on the plug-in estimator fit in $\mathcal{I}_1$ and the thresholds obtained in $\mathcal{I}_2$—to the observations in the validation cohort. Separating the creation of the classifier from the validation cohort allows the classifier to be used in data settings where the true class is unobserved.

Our end goal is to use the class labels in outcomes estimation. Thus, we force a single label per observation in a bootstrap procedure: for all n_{boot} resamples of the validation data, a single label, $\widehat{y}_{boot}$, is randomly selected with equal probability from all classes y where $\widehat{P}(y|x) \geq \widehat{t}_y$. For null sets, all class labels are considered with equal probability (eg, in settings where the classification algorithm has high average accuracy and there is little uncertainty, we may expect higher classification thresholds and thus null sets for difficult-to-classify observations). After a single label is randomly selected among those belonging to the sets $\widehat{C}_y$, outcome estimation can be performed based on $\widehat{y}_{boot}$. We refer to this approach as "weighted" labeling with bootstrap, or weighted bootstrap, because the probability of label assignment for a resampled observation is weighted by the set obtained prior to the resampling procedure. For example, in the case of $K = 3$, the set of possible label assignment weights is $\{0, 0.33, 0.5, 1\}$.

The bootstrap procedure allows the class labels assigned to a single observation to vary, thus incorporating classification uncertainty in both classification evaluation and outcomes estimation. Additionally, the resampling permits us to calculate confidence intervals for a variety of metrics of interest, including label coverage, counts of ambiguous label sets, conventional classification evaluation measures, and outcome measures. Because there is no sample mean for metrics that depend on a single class per observation (eg, classification accuracy), we report bootstrap percentile-based averages and confidence intervals.

Algorithm 1. Weighted labeling with bootstrap

---

1. For development data:

    a. Split development data into equal halves: $\mathcal{I}_1$ and $\mathcal{I}_2$

    b. Fit conditional probability estimator $P(y|x)$ on $\mathcal{I}_1$

    c. Apply $\widehat{P}(y|x)$ to $\mathcal{I}_2$ to obtain thresholds $\widehat{t}_y$

2. For validation data:

    a. Apply $\widehat{P}(y|x)$ and $\widehat{t}_y$ to obtain sets $\widehat{C}_y$ and calculate coverage

    b. Draw n_{boot} resamples:

       i. Select a single label $\widehat{y}_{boot}$ from each resampled observation's label set

       ii. Calculate classification performance measures based on $\widehat{y}_{boot}$

       iii. Estimate outcome of interest based on $\widehat{y}_{boot}$

    c. Report statistics of interest based on validation sample and validation bootstrap resamples

---

2.2 | Naive comparators: Most probable class, with and without resampling

As comparators, we implement a naive classification strategy where a single class label is assigned to each observation based on the most probable predicted class. Both naive comparators implement this classification approach. We use the same conditional probability estimator as in the weighted labeling approach, $P(y|x)$, but fit it on the entire development

cohort. The fitted classification algorithm is then applied to the validation cohort, and class labels are assigned based on the class with the highest predicted probability. These class labels are treated as known in subsequent outcomes estimation exercises. We use this approach, when applied in a single iteration of an analysis sample, as our standard practice comparator. A step beyond implementing this naive standard practice strategy is to incorporate bootstrap resampling. The classification algorithm fitting and prediction steps remain unchanged, but a bootstrap procedure is introduced for the classification performance evaluation and outcome estimation: the validation cohort is resampled and bootstrap-based confidence intervals may be calculated for classification and outcome measures. We include the naive bootstrap approach as a more nuanced method than the naive standard practice, but a less complex alternative to the weighted labeling bootstrap procedure introduced above. Depending on the real-world data context, the simpler, naive bootstrap may be preferable to the weighted labeling method.

2.3 | Classification performance measures

For a given $\widehat{P}(y|x)$, we assess classification performance via accuracy, sensitivity, specificity, and positive predictive value (see Table A1 for definitions). To calculate these measures, we tabulate the per-class proportions of true positives, true negatives, false positives, and false negatives. We then present what we call the macro-averaged versions of each measure by averaging across the classes. Because the split conformal inference-based LABEL classifier does not provide any coverage guarantees in this resampling setting, we empirically verify marginal coverage and examine estimated thresholds and ambiguous labels for each class. In our data application, we also plot the proportion of observations in a given stage by ordered predicted probability to assess calibration.

2.4 | Survival analysis

Survival outcomes are central to many strands of oncology research. Cancer stage indicates disease severity and is critical to assessing not only survival but also common health services research questions, such as examining spending outcomes and determining whether or not patients received appropriate treatment. Summarizing disease severity into a single variable simplifies both survival estimation and interpretability. To examine how our method compares to the naive approaches, we implement a common form of survival estimation, stratified Kaplan-Meier analysis. We select Kaplan-Meier estimation for simplicity and ease of exposition, but many alternative methods for survival outcome estimation could be used in practice, including standard parametric survival regression models, the Cox proportional hazards regression,[24] random survival forests,[25] and combinations of parametric, semiparametric, and nonparametric approaches in a stacked survival model.[26-28] To evaluate survival performance we examine bias and determine how the median survival time and 90 and 365-day survival probabilities based on predicted class compare to those based on observed class. We report average bias and 95% confidence intervals.

3 | SIMULATION STUDY

In this section, we describe a simulation study to compare the methods for classification and outcome estimation outlined above. The goal is to highlight settings where the relative performance of a naive, single iteration classification and outcome estimation approach ("standard practice") varies from a similar naive approach that incorporates a bootstrap procedure and from our proposed novel approach of weighted labeling with bootstrap, which leverages set-valued classification. Although we construct our simulation data to be similar to our real-world lung cancer data setting, the simulation study allows us to explore methodological performance under differing data scenarios. To reflect this, throughout the simulation study we refer to the classification outcome variable groups as "classes" rather than stage.

Our simulation sample is $N = 2000$ observations and we generate patient-level covariates $\mathbf{X}_i$ for each patient i, where $i = 1, ..., N$. To roughly approximate the sociodemographic and clinical characteristics of the lung cancer study, we simulate six continuous, seven binary, and two count variables; Table B2 provides additional details on covariate distributions. The class label outcome $Y = \{1, ..., K\}$, where $K = 3$, is generated from a multinomial distribution, $(Y_1, Y_2, Y_3) \sim \text{Multinom}(N, (p_1, p_2, p_3))$, where the class label Y_i is assigned based on the greatest class probability. Class

balance is approximately 37% class 1, 49% class 2, and 13% class 3. See Appendix B for additional details on the construction of the probabilities underlying Y.

To perform survival analysis, we use a Weibull distribution and simulate right censored event times: $T_i = (-\log(U_i)b \exp(-Y_i\beta_k))^{\frac{1}{a}}$, where T_i is simulated survival time influenced by time-independent covariate Y_i with effect parameter β_k, $U_i \sim$ Uniform$(0, 1)$, a is the Weibull shape parameter, and b the scale parameter. We set $a = 1$, $b = 90$, and Y_i is the class label simulated for the classification exercise. We set the length of observation time to 365 days and generate a variable to capture observed censored event times and an indicator if the event occurred or was censored. After simulating all variables, the data are stratified on the class label Y_i and divided into equally sized development and validation cohorts, each of size 1000. For the weighted labeling with bootstrap method, the development data are again stratified on Y_i and further divided into two halves of 500 observations each.

We implement three scenarios: (1) accurate and certain classification, (2) accurate and uncertain classification, and (3) inaccurate and uncertain classification. Table B2 compares the covariates used to fit the multinomial logistic regression for label classification in each scenario; scenarios differ only by which covariates are used to predict class label probabilities; all other simulation parameters are the same. We characterize the scenarios based on how accurately the classification regression predicts class labels with the given covariates, and based on the empirical distributions of the predicted probabilities (the uncertain scenarios have a higher proportion of predicted probabilities near 0.5). After fitting the conditional probability estimation algorithms and applying labels in the validation data, Kaplan-Meier survival estimation is performed in the validation data, stratified by observed and predicted class labels. We draw 500 bootstrap resamples for the bootstrap-based methods and we perform 1000 simulation repetitions for each scenario.

Across the simulation scenarios, we compare how each of the three methods performs in terms of classification and survival analysis. For classification performance we examine accuracy, sensitivity, specificity, and positive predictive value (PPV), as well as the underlying counts of true positive, true negatives, false positives, and false negatives. We present mean measures and bootstrap-based 95% confidence intervals where applicable. For all methods, we report coverage and 95% confidence intervals for each class. We also report the class-specific thresholds and ambiguity for the weighted labeling classifier (recall the naive methods do not incorporate threshold-based label assignment and only assign a single label per observation so there is no label ambiguity). Our primary survival analysis performance measure is bias. We examine the average bias across simulations and report percentile 95% confidence intervals around bias estimates.

3.1 | Simulation results

Class-specific coverage results across simulation scenarios are presented in Table 1. The weighted bootstrap method yields 90% coverage for all three classes across settings. Note that although we do not fill in null sets until the bootstrap procedure, because the number of null sets is low (described below), our classifier is still able to obtain the target coverage level. The naive classification method of assigning class labels based on the highest probability produces coverage ranging from <1% to 94% across all three classes, and coverage declines within each class from scenario 1 (accurate and certain class prediction) to scenario 2 (accurate and uncertain) to scenario 3 (inaccurate and uncertain). The naive classification methods do not target coverage, while class-specific target coverage is an explicit input of the LABEL classifier in the weighted bootstrap method. The two naive methods are identical at this stage because coverage is calculated in sample prior to the bootstrap procedure.

3.1.1 | Classification performance

The classifier thresholds in the weighted labeling with bootstrap method tend to decrease within class as uncertainty and inaccuracy are introduced across simulation scenarios (Table B3). Recall the thresholds are class-specific cutoff probabilities for determining if an observation receives a given class label. Class 2 is the most prevalent class (approximately 49% of the sample), class 1 is approximately 37% of the sample, and class 3 comprises approximately 13%. In scenario 1, the most accurate and certain setting, class 1 has the highest threshold, illustrating that the most prevalent class will not necessarily generate the highest label threshold.

Label set ambiguity provides a more complete picture as to how the classifier operates (Figure 2). As the accuracy and certainty declines across simulation scenarios, we see the number of null and single label sets decrease and the number of double and triple label sets increase. Examining ambiguous label sets by true class, we see where the distribution of label

TABLE 1 Coverage and Monte Carlo SEs

Method	Class 1	Class 2	Class 3
Scenario 1			
Naive labeling	0.94	0.87	0.41
	(0.01)	(0.01)	(0.00)
Weighted labeling	0.90	0.90	0.90
	(0.01)	(0.01)	(0.01)
Scenario 2			
Naive labeling	0.82	0.80	0.21
	(0.01)	(0.01)	(0.01)
Weighted labeling	0.90	0.90	0.91
	(0.01)	(0.01)	(0.01)
Scenario 3			
Naive labeling	0.54	0.62	0.00
	(0.02)	(0.02)	(0.00)
Weighted labeling	0.90	0.90	0.91
	(0.01)	(0.01)	(0.01)
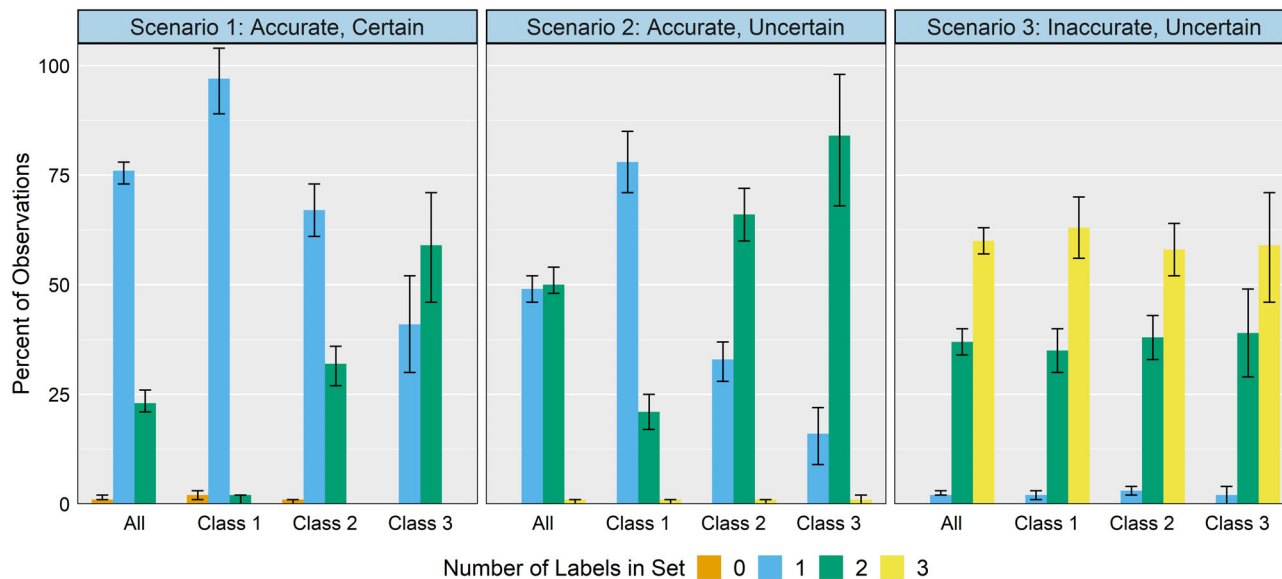
FIGURE 2 Simulation study label ambiguity: Share of sample by number of assigned labels in label set

sets differs by class. For example, in scenario 1 (accurate and certain), about 75% of all observations receive a single label set. But for class 1, over 95% of observations have a single label set, while in class 2 only 67% have a single label and in class 3 the proportion drops to 41%. Under scenario 2 (accurate and uncertain), 49% of all observations receive a single label set, but again this differs by true class: 78% of class 1 observations have a single label, 34% of class 2, and 21% of class 3. Under the inaccurate and uncertain scenario 3, the distribution becomes more even, with 2% of class 1 observations having a single label, 3% of class 2, and 2% of class 3. Label set ambiguity can be helpful as a diagnostic tool, particularly in the multiclass setting where uncertainty levels may vary across classes.

Figure 3 reports average classification accuracy, sensitivity, specificity, and PPV across simulation scenarios. Overall, the weighted labeling approach yields slightly poorer performance across metrics compared to the naive methods. Small differences can be seen for some metrics. For example, in scenario 2, weighted labeling has an accuracy of 77% while the
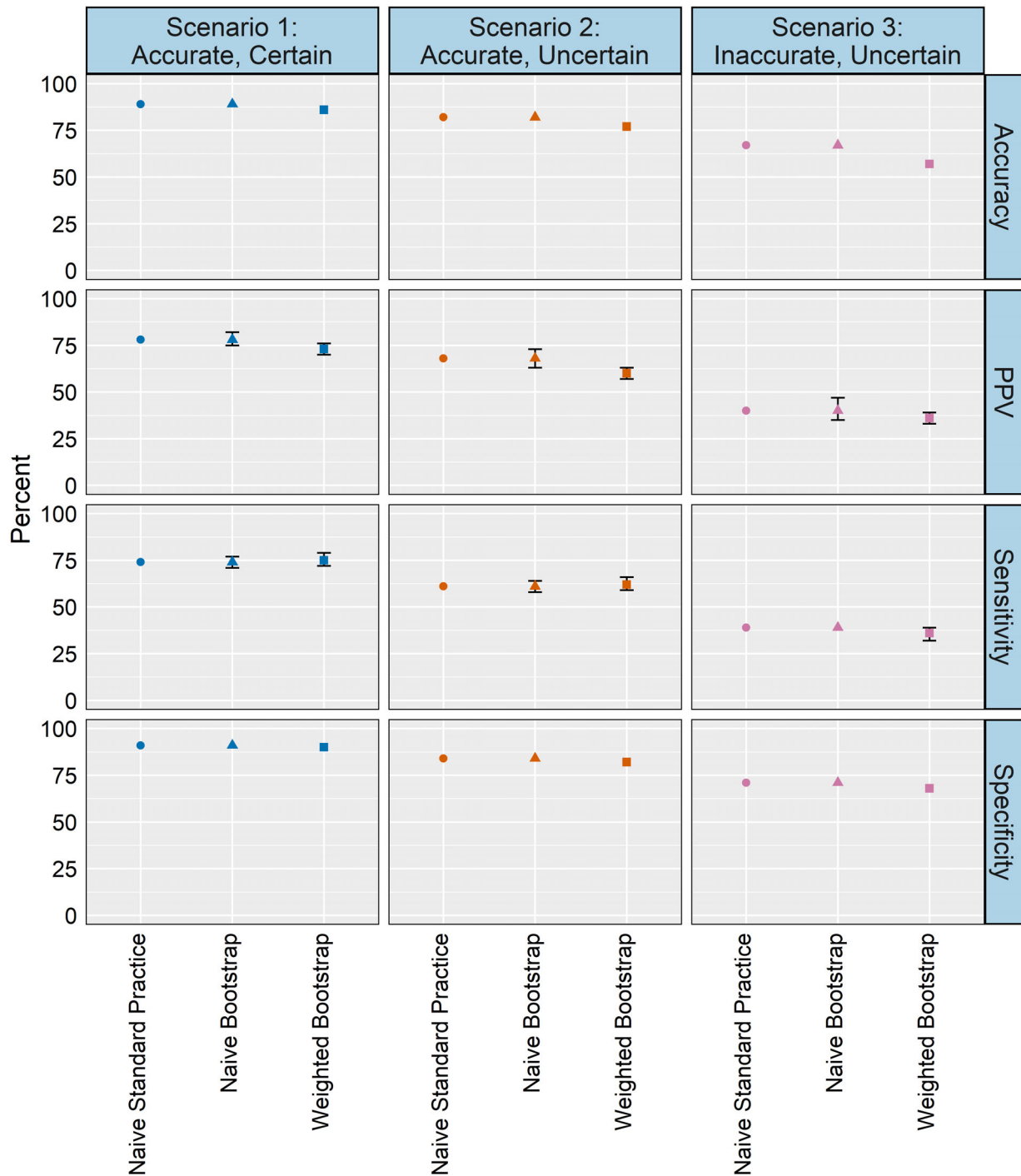
FIGURE 3 Simulation study classification discrimination: Average accuracy, sensitivity, specificity, and PPV (for visual clarity, 95% confidence intervals less than 0.05 are not displayed)

naive methods have an accuracy of 82%. Such differences may or may not be meaningful depending on the application. Table B4 reports class-specific classification accuracy. Results for class 1 are similar across the methods, but the weighted labeling performs worse compared to the naive methods in classes 2 and 3 by 4-7 percentage points.

Tables B5-B7 contain class-specific classification sensitivity, specificity, and PPV. The small differences between the naive methods and the weighted bootstrap are much more pronounced at the class-level. For example, the weighted bootstrap produces lower sensitivity for class 2 (by 12 percentage points in scenario 1, and about 25 percentage points in scenarios 2 and 3), but performs substantially better for class 3. In scenario 1, the weighted bootstrap sensitivity estimates are a 20 percentage point improvement over the naive methods, and an approximately 30 percentage point improvement
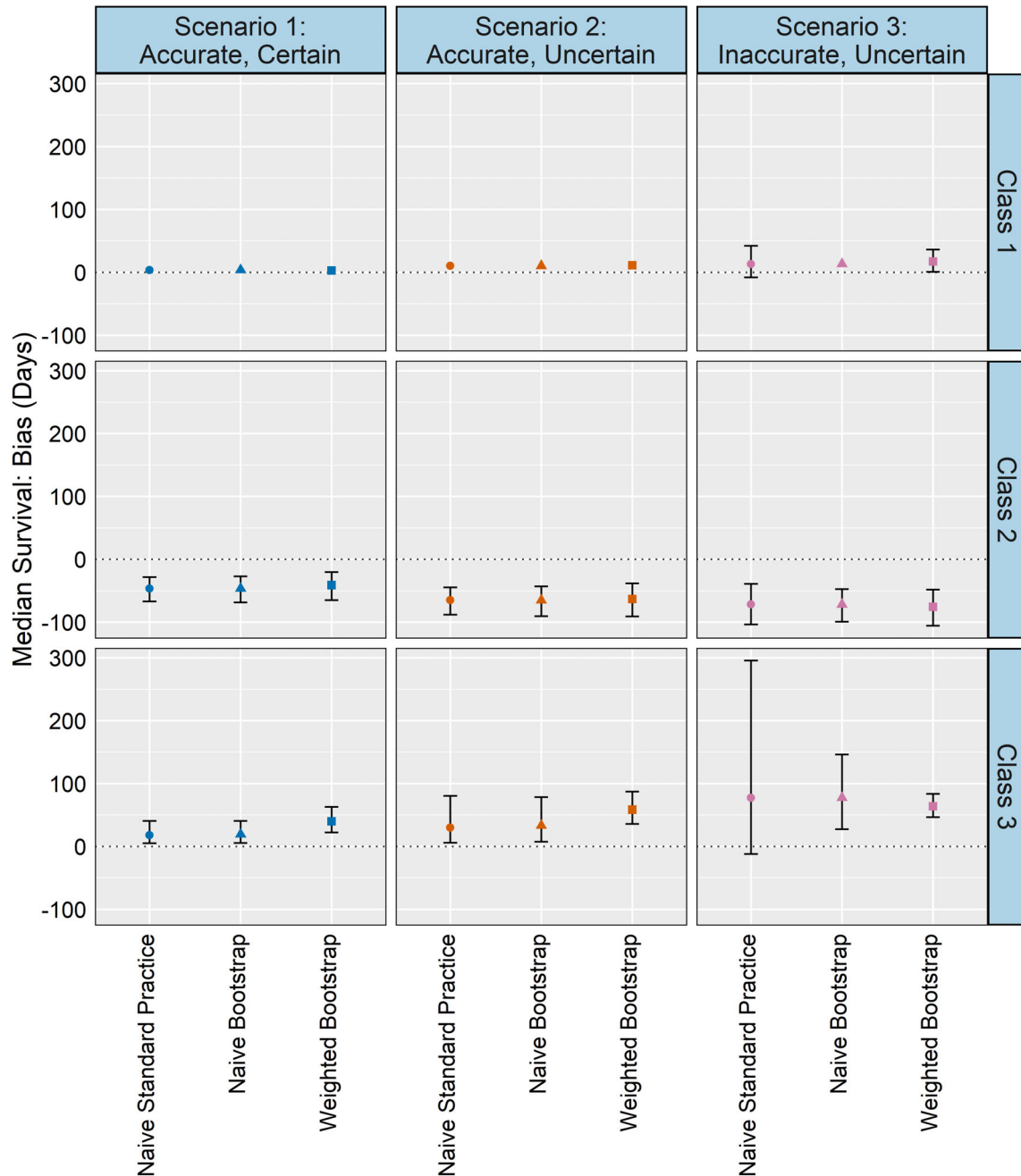
FIGURE 4 Simulation study median survival days bias (For visual clarity, 95% confidence intervals less than 31 days are not displayed)

in scenarios 2 and 3 (in scenario 3, the naive methods yield 0% sensitivity). Tables B8-B11 present true positive, true negative, false positive, and false negative counts by class across simulation scenarios. The conditional probability estimates for class 3 are generally low compared to classes 1 and 2, leading to few class 3 labels under the naive approaches. Across all settings, the weighted bootstrap method generates fewer true positives than the naive methods for classes 1 and 2, but yields a substantially larger number for class 3. The practical import of these class-based differences will vary by applied context, especially considering our primary outcome of interest here is survival.

3.1.2 │ Survival analysis

Bias for median survival is shown in Figure 4. Overall, the weighted bootstrap produces smaller bias about 40% of the time compared to the naive methods, but in each of these cases the confidence intervals are compatible with equal bias

across methods. The weighted bootstrap and naive methods have similarly sized confidence intervals across scenarios for classes 1 and 2. For class 3, the weighted bootstrap has a much smaller confidence interval than the naive methods, particularly so in scenario 3. The results for 90 and 365-day survival (Figures B1 and B2) run counter to what one might expect: rather than yielding smaller SEs or less biased average survival probabilities than the naive or weighted bootstrap methods, standard practice produces similar results and often larger confidence intervals, with the noted exception of class 3 in the inaccurate and uncertain scenario.

4 | DATA ANALYSIS

We use SEER cancer registry data linked with Medicare claims data as our real-world data application that motivated the development of these methods.[29] The SEER data provide information on all cancers diagnosed among individuals living in areas covered by SEER registries, including cancer stage.[30] SEER data are abstracted from medical records and contain validated staging information at the time of diagnosis, and thus are the source of our "gold standard" cancer stage labels: stage I/II, stage III, and stage IV. We combine stages I and II into a single label to accommodate sample size constraints and similarity of clinical outcomes.[30] Fee-for-service Medicare claims data contain detailed information on treatments received as well as health care visits and comorbidities. Medicare enrollment data provide information on patient age, race/ethnicity, vital status, and information on zip-code level measures of socioeconomic status.

Our study cohort includes individuals aged 65 and older who were enrolled in fee-for-service Medicare and were diagnosed with stage I-IV lung cancer between 2010 and 2013 who received chemotherapy within 6 months of diagnosis. We divide the data into two cohorts based on timing of diagnosis: a development cohort (2010-2011 diagnoses) and a validation cohort (2012-2013 diagnoses). In practice, a fixed classification algorithm is likely to be applied to individuals diagnosed in time periods successive to the training sample, thus we aim to approximate this likely real-world scenario. Table 2 shows the similarity between the two cohorts in terms of basic summary statistics; the similarity of our development and validation samples can be considered a "best case" scenario in terms of generalizability and prediction.

For stage classification, our input features are 94 variables derived from or linked to the Medicare claims data in the period 3 months before or after an individual receives their first lung cancer chemotherapy. Variables include patient

TABLE 2 Cohort summary statistics

Characteristic	Development cohort 2010-2011 diagnosis	Validation cohort 2012-2013 diagnosis
N	14 760	14 620
Age (mean)	72.1	71.9
Documented sex female (%)	45.4	46.8
Race/ethnicity (%)		
White	82.7	81.4
Black	8.8	8.9
Hispanic	4.2	4.4
Other	4.2	5.3
Region (%)		
Northeast	20.3	20.2
Midwest	13.6	13.2
West	34.2	34.5
South	31.9	32.2
Stage at diagnosis (%)		
I/II	15.1	14.8
III	34.0	33.1
IV	50.8	52.0

demographic characteristics, visits and hospitalizations, chemotherapy drugs, surgeries and procedures, radiation, comorbidities, and lung cancer anatomic site and malignancy diagnosis codes. A full list of features for this data set is described in Brooks et al.[9]

Prior work classifying lung cancer stage for patients receiving chemotherapy focuses on a binary split of early (stages I-III) vs late (stage IV).[8,9] Because it is unclear which single algorithm will perform best in the multiclass setting (stages I/II, stage III, and stage IV), we implement 7 algorithms that are multiclass versions of the most promising discrete binary prediction algorithms: main terms multinomial logistic regression;[31] penalized regressions (lasso, ridge, an elastic net with overall α penalty selected via internal cross-validation, and a balanced elastic net penalty set at 0.5),[32] generalized additive regression with cubic splines and a smoothing parameter set to 0.6;[33] random forests with node size 250 and 500 trees;[34] and gradient boosting with a maximum tree depth of 3, learning rate set to 1, and 2 fitting rounds.[35]

We examine patient survival to illustrate how our classification method can be used in practice with a common form of health outcomes estimation. Survival outcomes are estimated based on the number of days from first chemotherapy to death, and we follow patients for a one year period. For the bootstrap-based methods we report average bias, survival estimates, and percentile-based confidence intervals. For the naive standard practice approach we report bias, survival estimates, and Greenwood confidence intervals for the survival estimates. (Note that unlike in the simulation study, naive standard practice approach Monte Carlo-based SEs are not available for the data analysis.)

4.1 | Data analysis results

In the weighted labeling approach, all algorithms provide at least nominal coverage (90%) across all three stage classes. Under the naive method, for stage I/II, the average coverage is about 31%, stage III 60%, and stage IV about 84%. Table C15 contains the average coverage and 95% confidence intervals across bootstrap samples for each algorithm and method.

4.1.1 | Classification performance

The thresholds for the weighted labeling approach vary across algorithms and classes (Table C16). All of the algorithms produce the lowest thresholds for stage I/II and the highest thresholds for stage IV. The stage IV thresholds are similar across all algorithms (ranging from 0.31 to 0.40), but the random forests generates the lowest thresholds for stage I/II (0.00 vs 0.07-0.11) and stage III (0.07 vs 0.21-0.23).

None of the algorithms produce a null label set (multinomial logistic regression and random forests presented in Figure C3; results for other algorithms omitted due to similarity with multinomial logistic regression). Across all validation observations, random forests produces the fewest single label sets (about 24% of validation observations) and the most triple label sets (30%). In contrast, the multinomial logistic regression produces more single (35%) and double label sets (44%) across all classes. Examining ambiguity by true label class, we see the algorithms produce fewer single label sets for observations with a true class of stage I/II, reflecting the lower thresholds for this class. True stage IV is most likely to belong to a single label set, while true stages I/II and III are more likely to belong to a double label set, and all true classes have similar proportions of triple label sets.

To assess prediction calibration for each algorithm, we plot ordered predicted probabilities against the percent of observations belonging to a given stage (Figure C4). Recall the naive prediction methods use the entire development sample for algorithm fit, while the weighted bootstrap approach uses only half of the development sample for algorithm fitting, and the other half to set labeling thresholds. The average predicted probabilities across the distributions are nearly identical between the two methods for all stages and algorithms, with slight exceptions in gradient boosting and random forests. Overall, the algorithms are most poorly calibrated for stage I/II, show slightly better calibration for stage III, and perform best for stage IV.

Within the naive methods, the algorithms produce nearly identical classification performance results for several measures of discrimination (Figure 5). There are some small variations across algorithms within the weighted bootstrap method, but in general the classification performance of all algorithms—except the random forests—is within 1-2 percentage points. Tables C17-C20 present class-specific measures. The weighted bootstrap results are an improvement compared to the naive methods for stage I/II where sensitivity is higher (by 12-19 percentage points), and sensitivity and PPV estimates are higher for stage IV. However, the weighted bootstrap yields slightly lower accuracy across all three stage classes compared to the naive methods. Overall, in Figures C4 and 5, we see that measures of calibration and
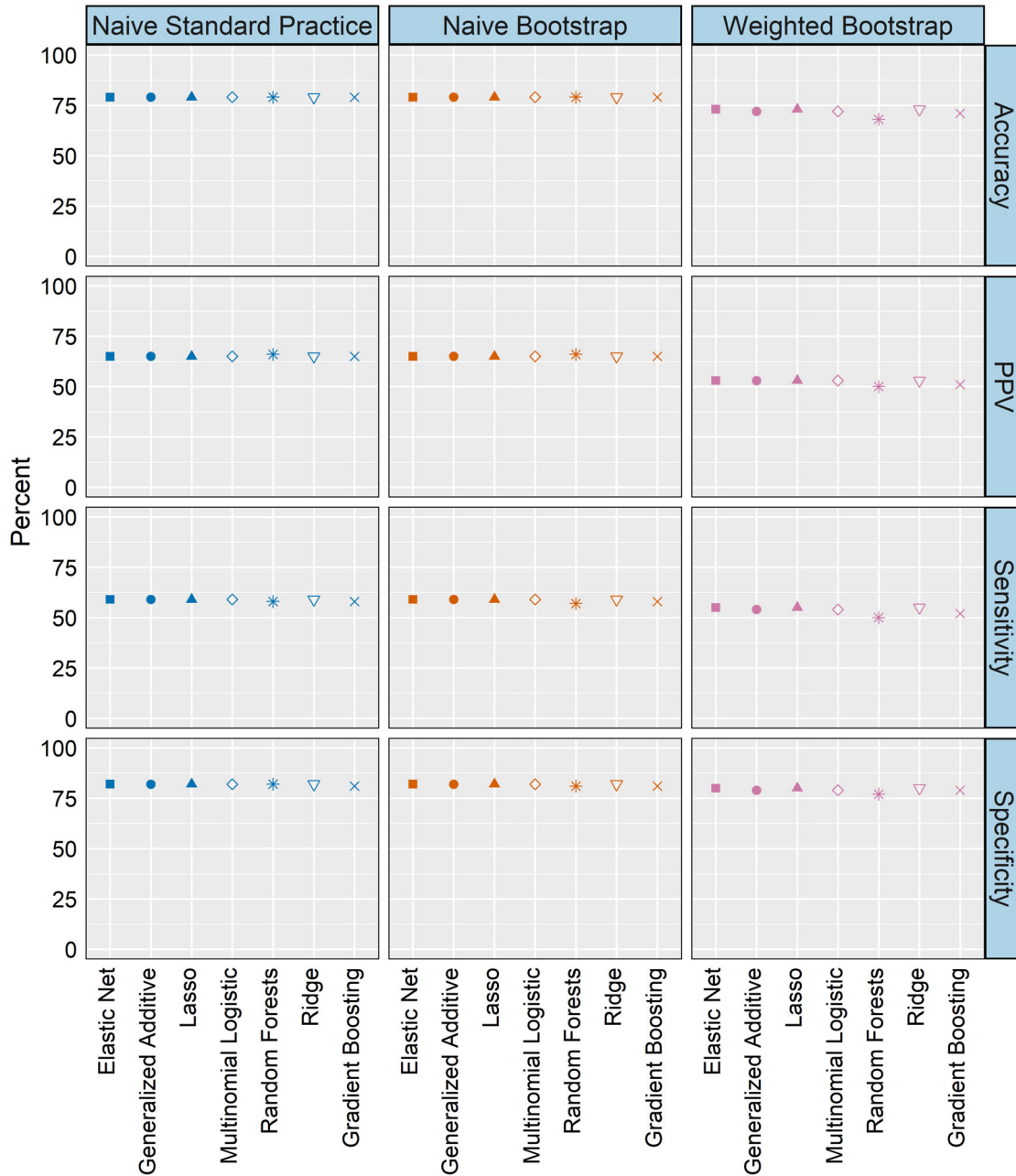
FIGURE 5 Data analysis classification discrimination: Average accuracy, sensitivity, specificity, and PPV (for visual clarity, 95% confidence intervals less than 0.05 are not displayed)

classification are similar, although in some cases there may be small differences, and performance by stage group is more variable.

4.1.2 | Survival analysis

For 90-day survival, the weighted bootstrap is close to 0% bias across all stages, whereas the naive methods show greater variation (Figure 6). The weighted bootstrap is approximately 5 percentage points better than the naive methods in stage I/II, about 2 percentage points better in stage III, and only about 2 percentage points worse than the naive methods for stage IV. The naive and weighted labeling bootstrap approaches yield similarly sized bootstrap-based confidence intervals. For 365-day survival, the weighted labeling method continues to generate the smallest bias for stages I/II and III, and in
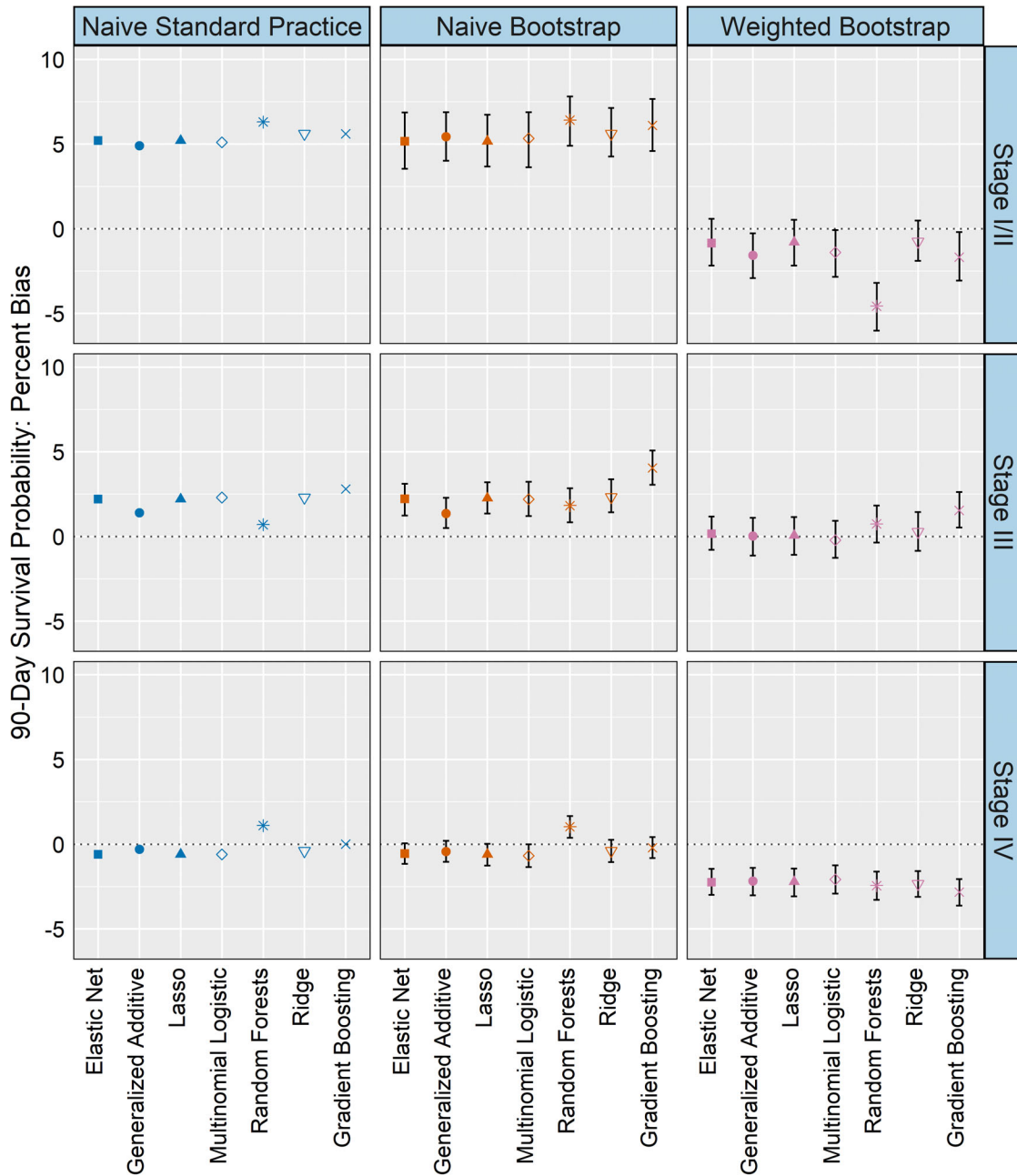
FIGURE 6 Data analysis: 90-day survival probability percent bias

some algorithms also produces the smallest bias in stage IV (Figure C5). The bias estimates are most different between the naive methods and the weighted label bootstrap for stage I/II, which is also the stage group that saw the greatest difference in classification results (Tables C17-C20). The observed median survival for stage I/II exceeds 1 year, so we do not estimate it here. For stage III, both naive methods tend to overestimate median survival compared to the true class, although they generate little bias for stage IV. The weighted labeling bootstrap method slightly underestimates median survival in both stage III and IV (Figure C6).

The naive standard practice and naive bootstrap methods generally produce similarly sized confidence intervals around the survival estimates, although there are some small variations in relative width across algorithms and class (see Figures C7-C9). Across all survival estimates, the weighted labeling bootstrap method produces similar results to the naive methods but does generate slightly narrower confidence intervals for the stage I/II 365-day survival estimates. As in the simulation study, we find the standard practice survival estimates do not necessarily yield smaller confidence intervals or

less biased average survival probabilities than the naive bootstrap method. Instead, they tend to produce similar results that vary by predicted class and the algorithm used to estimate conditional probabilities.

5 | DISCUSSION

In this article, we studied conveying uncertainty in applied classification settings, and proposed a procedure leveraging set-valued classification, split conformal inference, and resampling. Our proposed method uses bootstrap resampling from the sets of plausible labels generated in the classification step, and then performs outcomes estimation based on the selected labels. In our real-world data example, we developed fixed multiclass prediction algorithms for labeling lung cancer stage. The weighted labeling procedure yielded the smallest bias for survival estimates in stages I/II and stage III.

Our development and validation samples had similar observed characteristics and were drawn from sequential time periods: development cohort patients were diagnosed with lung cancer from 2010 to 2011 and validation cohort patients in 2012-2013. As such, the naive bootstrap approach may in practice be preferable to the weighted labeling bootstrap in such scenarios due to the simplicity of implementation without dramatic losses in performance. However, had our validation sample been drawn from a very different population—for example, patients diagnosed in 2019-2020—we might find much higher levels of uncertainty and lower levels of prediction accuracy due to changes in treatment patterns and patient characteristics. In this case, the weighted labeling approach may provide a more nuanced picture of the label uncertainty and impact on outcomes estimation by generating label sets for each observation.

Simulations show that our method outperforms the naive methods in terms of bias for some classes and scenarios, and for others yields similar survival estimates and confidence intervals. The simulation study also illustrates how label uncertainty can vary across classes and be translated into poor outcomes estimation performance; for example, class 3 yielded a higher number of ambiguous label sets and the greatest amount of bias across all simulation scenarios. Further, we note that our method helps highlight the issues with making predictions in uncertain or inaccurate data scenarios: reporting label ambiguity in addition to multiple performance metrics emphasizes poor prediction characteristics and algorithms that should not be used in downstream analyses.

The goal of this study was to propose an approach for characterizing uncertainty from a prediction exercise and demonstrate how this uncertainty can be incorporated in downstream survival analysis. As noted in the text, while we implemented a simple Kaplan-Meier analysis, there are more sophisticated time-to-event estimation approaches that could be deployed. Similarly, there are a range of other outcomes that may be of interest, including adjusted quality of care measures or cost-effectiveness estimates for oncology treatments.

In summary, the weighted labeling with bootstrap method is one approach for incorporating uncertainty in applied classification and estimation problems. The proposed naive bootstrap procedure is also an improvement over simply ignoring prediction uncertainty. Depending on the data setting, the naive bootstrap may be implemented as an uncomplicated alternative to the weighted labeling method. As classification and risk prediction algorithms become more commonplace in medical and health services settings, we must think beyond prediction evaluation and implement tools to effectively communicate uncertainty.

State of California Department of Public Health, the National Cancer Institute, and the Centers for Disease Control and Prevention or their Contractors and Subcontractors is not intended nor should be inferred.

CONFLICT OF INTEREST
The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT
The data that support the findings of this study are available from the SEER-Medicare database, which is a conjoint effort by the National Cancer Institute; the Office of Research, Development, and Information, CMS; Information Management Services (IMS), Inc.; and the Surveillance, Epidemiology, and End Results (SEER) Program tumor registries. Restrictions apply to the availability of these data, which were used under Data Use Agreement license for this study. Supporting code associated with this article may be found at: https://github.com/sl-bergquist/cancer_classification

ORCID
Savannah Bergquist https://orcid.org/0000-0001-6495-0724
Sherri Rose https://orcid.org/0000-0002-9076-8472

REFERENCES
1. Degtiar I, Rose S. A review of generalizability and transportability; 2021. arXiv:2102.11904.
2. Steingrimsson JA, Gatsonis C, Dahabreh IJ. Transporting a prediction model for use in a new target population; 2021. arXiv:2101.11182.
3. Keating NL, Jhatakia S, Brooks GA, et al. Association of participation in the oncology care model with medicare payments, utilization, care delivery, and quality outcomes. *JAMA*. 2021;326(18):1829-1839. doi:10.1001/jama.2021.17642
4. Newcomer LN, Gould B, Page RD, Donelan SA, Perkins M. Changing physician incentives for affordable, quality cancer care: results of an episode payment model. *J Oncol Pract*. 2014;10(5):322-326. doi:10.1200/JOP.2014.001488
5. Hassett MJ, Uno H, Cronin AM, Carroll NM, Hornbrook MC, Ritzwoller D. Detecting lung and colorectal cancer recurrence using structured clinical/administrative data to enable outcomes research and population health management. *Med Care*. 2017;55(12):e88-e98. doi:10.1097/MLR.0000000000000404
6. Ritzwoller DP, Hassett MJ, Uno H, et al. Development, validation, and dissemination of a breast cancer recurrence detection and timing informatics algorithm. *J Natl Cancer Inst*. 2018;110(3):273-281. doi:10.1093/jnci/djx200
7. Esposito DB, Russo L, Oksen D, et al. Development of predictive models to identify advanced-stage cancer patients in a US healthcare claims database. *Cancer Epidemiol*. 2019;61:30-37. doi:10.1016/j.canep.2019.05.006
8. Bergquist SL, Brooks GA, Keating NL, Landrum MB, Rose S. Classifying lung cancer severity with ensemble machine learning in health care claims data. In: Doshi-Velez F, Fackler J, Kale D, Ranganath R, Wallace B, Wiens J, eds. *Proceedings of the 2nd Machine Learning for Healthcare Conference, Proceedings of Machine Learning Research*. Vol 68. Boston, MA: PMLR; 2017:25-38.
9. Brooks GA, Bergquist SL, Landrum MB, Rose S, Keating NL. Classifying stage IV lung cancer from health care claims: a comparison of multiple analytic approaches. *JCO Clin Cancer Inform*. 2019;3(3):1-19. doi:10.1200/CCI.18.00156
10. Wang S, McCormick TH, Leek JT. Post-prediction inference. bioRxiv; 2020. 10.1101/2020.01.21.914002
11. Ogburn EL, Rudolph KE, Morello-Frosch R, KA, Casey JA. A warning about using predicted values from regression models for epidemiologic inquiry. *Am J Epidemiol*. 2020;190:1142-1147.
12. Meijerink L, Cina G, Tonutti M. Uncertainty estimation for classification and risk prediction on medical tabular data; 2020. arXiv:2004.05824.
13. Raghu M, Blumer K, Sayres R, et al. Direct uncertainty prediction for medical second opinions. In: Chaudhuri K, Salakhutdinov R, eds. *Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research*. Vol 97. Cambridge, MA: PMLR; 2019:5281-5290.
14. Vovk V, Gammerman A, Shafer G. *Algorithmic Learning in a Random World*. New York, NY: Springer; 2005.
15. Papadopoulos H. Inductive conformal prediction: Theory and application to neural networks; 2008; Citeseer.
16. Vovk V. Conditional validity of inductive conformal predictors. In: Hoi SCH, Buntine W, eds. *Proceedings of the Asian Conference on Machine Learning, Proceedings of Machine Learning Research*. Vol 25. Singapore: PMLR, Singapore Management University; 2012:475-490.
17. Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L. Distribution-free predictive inference for regression. *J Am Stat Assoc*. 2018;113(523):1094-1111. doi:10.1080/01621459.2017.1307116
18. Nestor B, McDermott MBA, Boag W, et al. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. In: Doshi-Velez F, Fackler J, Jung K, et al., eds. *Proceedings of the 4th Machine Learning for Healthcare Conference, Proceedings of Machine Learning Research*. Vol 106. Ann Arbor, Michigan: PMLR; 2019:381-405.
19. Shafer G, Vovk V. A tutorial on conformal prediction. *J Mach Learn Res*. 2008;9(12):371-421.
20. Sadinle M, Lei J, Wasserman L. Least ambiguous set-valued classifiers with bounded error levels. *J Am Stat Assoc*. 2019;114(525):223-234. doi:10.1080/01621459.2017.1395341
21. Vovk V. Cross-conformal predictors. *Ann Math Artif Intell*. 2015;74(1):9-28. doi:10.1007/s10472-013-9368-4

22. Vovk V, Nouretdinov I, Manokhin V, Gammerman A. Cross-conformal predictive distributions. In: Gammerman A, Vovk V, Luo Z, Smirnov E, Peeters R, eds. *Proceedings of the Seventh Workshop on Conformal and Probabilistic Prediction and Applications, Proceedings of Machine Learning Research.* Vol 91. Cambridge, MA: PMLR; 2018:37-51.

23. Barber RF, Candes EJ, Ramdas A, Tibshirani RJ. Predictive inference with the jackknife+ 2019. arXiv:1905.02928.

24. Cox DR. Regression models and life-tables. *J Royal Stat Soc Ser B (Methodol)*. 1972;34(2):187-220.

25. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat*. 2008;2(3):841-860. doi:10.1214/08-AOAS169

26. Hothorn T, Bühlmann P, Dudoit S, Molinaro A, Van Der Laan MJ. Survival ensembles. *Biostatistics*. 2005;7(3):355-373. doi:10.1093/biostatistics/kxj011

27. Wey A, Connett J, Rudser K. Combining parametric, semi-parametric, and non-parametric survival models with stacked survival models. *Biostatistics*. 2015;16(3):537-549. doi:10.1093/biostatistics/kxv001

28. Golmakani MK, Polley EC. Super learner for survival data prediction [published online ahead of print, 2020 Feb 22]. *Int J Biostat*. 2020. doi:10.1515/ijb-2019-0065

29. Enewold L, Parsons H, Zhao L, et al. Updated overview of the SEER-medicare data: enhanced content and applications. *JNCI Monogr*. 2020;2020(55):3-13. doi:10.1093/jncimonographs/lgz029

30. Howlader N, Noone AM, Krapcho M, et al. SEER cancer statistics review 1975-2017. Report, National Cancer Institute; 2019.

31. Venables WN, Ripley BD. *Modern Applied Statistics with S*. 4th ed. New York, NY: Springer; 2002.

32. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1-22.

33. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J Royal Stat Soc B*. 2011;73(1):3-36.

34. Liaw A, Wiener M. Classification and regression by random forest. *R News*. 2002;2(3):18-22.

35. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. New York, NY, USA: Association for Computing Machinery, 2016:785-794. https://doi.org/10.1145/2939672.2939785

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.