Classifying Lung Cancer Severity with Ensemble Machine Learning in Health Care Claims Data

Savannah L. Bergquist BERGQUIST@G.HARVARD.EDU
Department of Health Care Policy
Harvard Medical School, Boston, MA, USA
Gabriel A. Brooks GABRIEL.A.BROOKS@HITCHCOCK.ORG
Department of Medicine
Dartmouth-Hitchcock Medical Center, Lebanon, NH, USA
Nancy L. Keating KEATING@HCP.MED.HARVARD.EDU
Department of Health Care Policy
Harvard Medical School and
Division of General Internal Medicine
Brigham and Women's Hospital, Boston, MA, USA
Mary Beth Landrum LANDRUM@HCP.MED.HARVARD.EDU
Department of Health Care Policy
Harvard Medical School, Boston, MA, USA
Sherri Rose ROSE@HCP.MED.HARVARD.EDU
Department of Health Care Policy
Harvard Medical School, Boston, MA, USA

Abstract

Research in oncology quality of care and health outcomes has been limited by the difficulty of identifying cancer stage in health care claims data. Using linked cancer registry and Medicare claims data, we develop a tool for classifying lung cancer patients receiving chemotherapy into early vs. late stage cancer by (*i*) deploying ensemble machine learning for prediction, (*ii*) establishing a set of classification rules for the predicted probabilities, and (*iii*) considering an augmented set of administrative claims data. We find our ensemble machine learning algorithm with a classification rule defined by the median substantially outperforms an existing clinical decision tree for this problem, yielding full sample performance of 93% sensitivity, 92% specificity, and 93% accuracy. This work has the potential for broad applicability as provider organizations, payers, and policy makers seek to measure quality and outcomes of cancer care and improve on risk adjustment methods.

1. Introduction

In the United States, it is estimated that 222,500 individuals will be diagnosed with lung cancer in 2017, and 155,870 will die of lung cancer (Siegel et al., 2017). Stage at lung cancer diagnosis is the most important factor associated with survival; the 5-year relative survival is 55% for those with localized disease compared with 4% for those with metastases at the time of diagnosis (Howlader et al., 2016). Historically, most lung cancers are diagnosed at late stages, when the chance for a cure is lower, although this may change with recent recommendations for lung cancer screening among current or former smokers (Moyer, 2014).

Payers and other healthcare organizations are increasingly using administrative data to measure quality of care and patient outcomes. Oncology care is one area that has posed challenges to large-scale quality measurement due to the crucial importance of understanding clinical stage in assessing outcomes and the quality of care delivered. Identifying cancer stage from administrative data has numerous challenges, and studies that used claims-based algorithms to identify metastatic cancers, recurrence, or progression have not produced tools with consistently high sensitivity and specificity (Hassett et al., 2014; Chawla et al., 2014; Warren and Yabroff, 2015; Nordstrom et al., 2016). While some of the best performing algorithms have been decision trees that rely on secondary malignancy ICD9 codes and chemotherapy agents (Whyte et al., 2015; Nordstrom et al., 2012), published research in claims data has demonstrated an inability to achieve sensitivity and specificity in the full sample above 80% simultaneously. Although administrative claims data have become a commonly used source of "big data," the absence of reliable claims-based classification algorithms is a substantial barrier to conducting lung cancer outcomes research at a population level.

A decision tree based on clinical guidelines for care (NCCN, 2017) has been developed to predict early-stage lung cancer using cancer registry data linked with Medicare administrative data (Brooks et al., 2017). This clinical tree had poor performance in identifying patients with early-stage cancer, particularly with respect to sensitivity. In this paper, we aim to improve upon the clinical tree, and develop a tool for classifying lung cancer severity by (*i*) deploying ensemble machine learning for prediction, (*ii*) establishing a set of classification rules for the predicted probabilities, and (*iii*) using an augmented set of administrative claims data.

2. Study Cohort

Our study cohort contains detailed information on Medicare beneficiaries with lung cancer, combining data from the Surveillance, Epidemiology and End Results (SEER) cancer registry program and Medicare claims data (Potosky et al., 1993). SEER data provide a "gold standard" for assessing algorithm performance for staging because these data are abstracted from hospital medical records and contain reliable tumor morphology and staging information at the time of diagnosis. Additionally, fee-for-service Medicare claims data provide a rich set of variables for developing a classification algorithm and studying health outcomes.

2.1 Cohort Selection

Our data included cancers diagnosed in 2010-2011 linked with Medicare claims from 2009-2012 who received at least one dose of infused or oral chemotherapy within six months of diagnosis. We combined the SEER registry data with the Medical Provider Analysis and Review (Part A short inpatient stay, long inpatient stay, and skilled nursing facility bills with one record per admission), National Claims History (Part B claims for non-institutional providers), Durable Medical Equipment (final action claims), outpatient claims (Part B claims from institutional outpatient providers), and Part D (prescription drug events) files. The National Claims History and Durable Medical Equipment files were combined at the observation level, and the outpatient files at the claim level; Part D files were included if there were any National Claims History or outpatient records 59 days prior to the lung cancer diagnosis or including the lung cancer diagnosis. Additionally, we combined our cohort file with information on patient comorbidities from the Medicare Chronic Conditions Warehouse and on census tract-level variables from SEER-Medicare. After identifying 74,630 patients with known month and year of diagnosis, we excluded 4,522 patients with unknown stage and 26,323 who were not continuously enrolled in Parts A and B of fee-for-service Medicare during the month of diagnosis and following six months. Of these 42,069, we identified 14,743 patients who were treated with chemotherapy within six months of diagnosis, had a lung cancer diagnosis on the chemotherapy claim, and had census tract-level information.

2.2 Feature Choices

To create our primary outcome variable, we grouped stages I-III into early stage and treated stage IV as late stage. Given differences in health outcomes for stage IV patients, this is often a preferred grouping. The clinical tree algorithm, developed in earlier work on a single split sample, is a decision tree with seven nodes that map to the binary outcome, described in Algorithm 1 (Brooks et al., 2017).

Algorithm 1: *Clinical Tree.*

```
⋆ For each observation i[§]:
  ⋆ if no lung cancer-specific chemotherapy, then early stage;
      ⋆ else if advanced non-small cell lung cancer chemotherapy[†]
        or stereotactic cranial radiation[†], then late stage;
      ⋆ else if lung resection surgery[‡], then early stage;
      ⋆ else if radiation[‖], then early stage;
      ⋆ else if small-cell chemotherapy agents and platinums only[†], then late stage;
      ⋆ else if targeted agents[†], then late stage;
      ⋆ else late stage.
```

[§] who received any chemotherapy within 6 months of diagnosis
[†] within 3 months of initial lung cancer chemotherapy
[‡] lobectomy, pneumonectomy, or segmental resection in 3 months before initial lung cancer chemotherapy
[‖] 20 or more fractions beginning not more than 7 days before initial lung chemotherapy

Since the clinical tree algorithm is based on clinical guidelines for lung cancer care, it relies only on a small targeted set of clinical variables. It is of interest to compare how this set of limited variables – selected through investigator knowledge and national cancer treatment guidelines – performs relative to a larger more comprehensive set of variables. Thus, to augment the set of seven clinical variables in an effort to improve classification performance, we included additional variables that are readily available to the Centers for Medicare and Medicaid Services in administrative claims data. Specifically, we considered a broad set of features categorized into seven groups: demographic (25), visits and hospitalizations (10), chemotherapy drugs (30), surgery and procedures (4), radiation (19), comorbidities (14), and lung cancer anatomic site codes and secondary malignancy diagnosis codes (13). This last group of variables, ICD9 diagnosis codes for lung cancer anatomic site and secondary malignancies, has been shown to be unreliable in previous population-based studies of lung cancer (Cooper et al., 1999; Nordstrom et al., 2012; Chawla et al., 2014; Warren and Yabroff, 2015; Whyte et al., 2015). Thus, we will consider three sets of variables, clinical algorithm only (7), clinical augmented with administrative claims except diagnoses codes (102), and all (115). See Appendix A, Table 4 for a description of each variable. Select demographic information is summarized in Table 1.

Table 1: Demographic Summary by Stage, $n = 14,743$

	Early	Late
Age, mean years (sd)	72 (8)	72 (8)
Male, %	54	55
White, %	83	83
Census Tract Below Poverty, %	12	11
Census Tract Non-High School Graduate, %	21	20

3. Methods

The goal of this analysis is to improve upon the clinical tree algorithm for classifying lung cancer patients into early and late stage based on SEER-Medicare data. Previous studies implementing machine learning methods to predict or classify different types of cancers have used neural nets, support vector machines (SVMs), decision trees, and logistic regression with varying degrees of success (Konstantina et al., 2015). Thus, in this setting, it is particularly unclear which single algorithm or set of variables will have optimal performance. Therefore, we deploy the super learner ensemble framework to build our prediction function (van der Laan et al., 2007) and consider multiple variable sets. The super learner yields an optimal weighted combination of candidate algorithms according to a specified loss function.

3.1 Data, Model, and Parameter

We now introduce formal notation for this applied problem. Our data structure is defined as $O = (Y, C)$, where Y is our binary outcome for lung cancer severity, with $Y = 1$ indicating early stage, and C our vector of covariates. This vector of covariates can be broken into three mutually exclusive subsets $C = (C_1, C_2, C_3)$, with C_1 including only the seven clinical variables used by the clinical tree from Algorithm 1, C_2 containing 95 demographic, claims, treatment, and comorbidity variables, and the 13 lung cancer type and secondary malignancy diagnosis codes comprising C_3. We write $C_1 \cup C_2$ as C_{12} to represent the "clinical augmented with administrative claims except diagnoses codes" set of variables and $C_1 \cup C_2 \cup C_3$ as C_{123} to represent the set of "all" variables, including the potentially unreliable ICD9 codes for lung cancer type and secondary malignancies. (Note that $C_{123} = C$, but we use C_{123} in places to be explicit.) We also write $C_{(\cdot)}$ when the set of covariates may be any of C_1, C_{12}, C_{123}, etc.

We consider a nonparametric model $\mathcal{M}$ that is the set of possible probability distributions, and describe the observational unit O as being drawn from true probability distribution P_0, where subscript 0 indicates the unknown truth. Succinctly, $O \sim P_0$ and $P_0 \in \mathcal{M}$.

Our nonparametric model assumes that our data are i.i.d., but does not impose additional functional form assumptions on the generation of Y, for example. Our parameter of interest for the prediction problem is $\Psi(P_0) = P_0(Y = 1 \mid C_{(\cdot)})$, which can also be written $\Psi(P_0) = \arg\min_{\Psi(P)} E_0 L(O, \Psi(P))$, where $E_0 L(O, \Psi(P))$ is the expected loss. Given our binary Y, both the squared error loss and negative log loss functions target the same parameter $\Psi(P_0) = P_0(Y = 1 \mid C_{(\cdot)})$. Thus, we use $L(O, \Psi(P)) = (Y - \hat{\Psi}(P))^2$, where $\hat{\Psi}(P)$ is any estimator of $\Psi(P_0)$, and we seek to minimize the expected loss when building our prediction function. We describe additional evaluation metrics for the overall tool in Section 3.4.

3.2 Estimation: Ensembling for Prediction

We describe the ensemble approach super learner (van der Laan et al., 2007) in Algorithm 2.

```
Algorithm 2: Super Learner.
⋆ For each algorithm $k$:
    ⋆ Perform $V$-fold cross validation, obtaining cross-validated predicted values $Z_k$;
    ⋆ Fit on full data $O$, obtaining $\hat{\Psi}(P)_k$;
⋆ Index a proposed family of convex combinations of the $k$ algorithms by $\alpha$;
⋆ Select $\hat{\alpha}$ to minimize $E_0 L(O, \Psi(P))$, which can be shown is solved by estimating:
        $\text{logit}(\hat{P}(Y = 1|Z)) = \alpha_1 Z_1 + .... + \alpha_k Z_k$;
⋆ Save $\hat{\Psi}(P)_{SL}$, the final estimator of $\Psi(P_0) = P_0(Y = 1 \mid C)$, constructed as:
        $\hat{\Psi}(P)_{SL} = \hat{\alpha}_1 \hat{\Psi}(P)_1 + \ldots + \hat{\alpha}_K \hat{\Psi}(P)_K$.
Note: The entire super learner algorithm above is itself externally cross-validated to obtained cross-validated performance metrics.
```

Our implementation of super learner considered eight algorithms three times, once for each of the variable sets C_1, C_{12}, and C_{123}, as well as the clinical tree in Algorithm 1, which, by definition, only uses C_1. Thus, we consider a convex combination of a total of $K = 25$ algorithms with our super learner forming a separate 26th algorithm. The eight algorithms were: (a) random forest with a node size of 250 and 500 trees; (b) neural net with two units in the hidden layer; (c) main terms logistic regression (GLM); (d) generalized additive model; (e) lasso penalized regression with λ chosen via internal cross-validation; (f) ridge regression with λ chosen via internal cross-validation; (g) balanced elastic net regression with $\alpha = 0.5$ and λ chosen via internal cross-validation; and (h) SVM with a cost parameter of 1 and a Gaussian kernel width γ parameter of $1/\text{length}(C_{(\cdot)})$. This specific implementation is also visualized in Figure 1. The analysis was performed using 10-fold cross-validation in `R` version 3.3.2 on an Oracle Sun Server X4-4 with 60 cores and 1.5TB of RAM with Linux software relying on the `SuperLearner` package (Polley et al., 2016).

3.3 Classification Rules

We establish two main thresholding rules for classifying probabilities into stage categories: (I) assignment based on the most likely class, and (II) assignment based on the percentile distribution of predicted probabilities. Within the first rule based on fixed probabilities, we explore thresholds of 50% ± 10 percentage points. A rule based on the median may provide better performance than a rule based on most likely class when the population is approximately balanced between the two outcomes and the algorithm provides strong

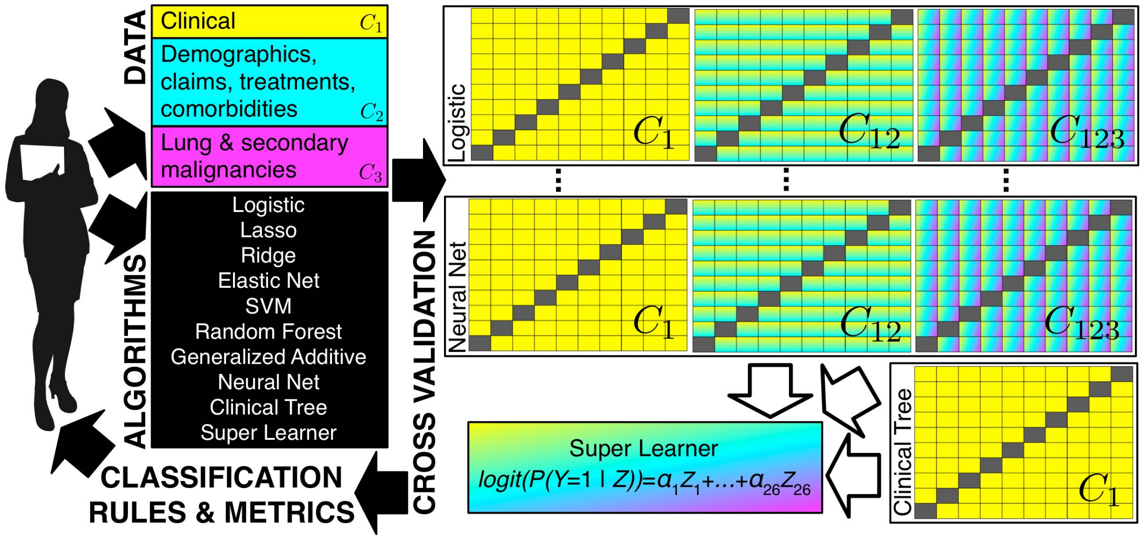

Figure 1: Flowchart for Lung Cancer Severity Classification Tool

discrimination. In our sample, the outcome group sizes are well balanced (49% had early stage cancer), but, in general, the proportion of patients with early stage cancer varies by cancer type and population. In unbalanced samples, or in cases where the algorithm has poor discrimination between the outcomes, a rule based on the empirical percentile distribution may be biased, while a rule based on most likely stage may provide better performance. To explore the potential for misclassification, we test five total percentile based rules: 30^{th}, 40^{th}, 50^{th}, 60^{th}, and 70^{th}. Our classification rules are implemented and evaluated as part of the last step in our classification tool shown in Figure 1.

3.4 Evaluation Metrics

For prediction, we evaluate each of our 26 algorithms according to cross-validated mean squared error (MSE $= (Y_i - Z_{k,i})^2$), cross-validated $R^2 = 1 - (\sum_i (Y_i - Z_{k,i})^2)/(\sum_i (Y_i - \bar{Y}_i)^2)$, and a cross-validated relative efficiency, RE = CV MSE$_k$/CV MSE$_{SL}$. The best prediction function is selected based on lowest cross-validated MSE, but we do evaluate non-selected prediction functions in the next step for completeness. For classification, we then consider sensitivity, specificity, accuracy (defined as the proportion of true positives *and* true negatives), and area under the curve (AUC) for each rule specified. The published literature for cancer staging almost exclusively evaluates sensitivity, specificity, accuracy, and AUC in the full sample, and the goal of achieving sensitivity and specificity $\geq 80\%$ is a metric for the full data. However, cross-validated metrics may be preferable to assess issues such as overfitting. Thus, we consider both full sample and cross-validated versions of these metrics. We establish the best performing classification rule to be the one that obtains the highest cross-validated AUC. Additionally, we plot the proportion of observed early stage lung cancer observation by ordered final predicted probability for our selected algorithm. Our metrics form the final component of the classification tool in Figure 1.

4. Results

The super learner prediction function contained five algorithms with nonzero weights:

$$\begin{aligned}\hat{\Psi}(P)_{SL} &= 0.43\hat{\Psi}(P)_{\texttt{rf}_{C_{123}}} + 0.38\hat{\Psi}(P)_{\texttt{gam}_{C_{123}}} \\ &\quad +0.13\hat{\Psi}(P)_{\texttt{lasso}_{C_{123}}} + 0.03\hat{\Psi}(P)_{\texttt{svm}_{C_{123}}} + 0.03\hat{\Psi}(P)_{\texttt{rf}_{C_{12}}},\end{aligned}$$

where `rf` is random forest, `gam` is the generalized additive model, `lasso` is the lasso regression, `svm` is the SVM, and $C_{(\cdot)}$ subscripts indicate the variable set used. This super learner algorithm for predicting early stage cancer improved substantially upon the clinical tree algorithm with a cross-validated R^2 of 0.405 and a cross-validated MSE of 0.149. The clinical tree had a relative efficiency of only 0.30, and its negative cross-validated R^2 indicates that the mean probability performs better than the predicted probabilities generated by the clinical tree. While the super learner had the overall best performance based on MSE, there were several individual algorithms, all using variable set C_{123} that had relative efficiencies of 94-98%. (See Table 2 for algorithm performance ranked by relative efficiency.)

Table 2: Ranked Algorithm Performance

Algorithm	CV R^2	CV MSE	RE
Super Learner	0.405	0.149	1.00
GAM: C_{123}	0.396	0.151	0.98
Lasso: C_{123}	0.395	0.151	0.98
Ridge: C_{123}	0.395	0.151	0.98
Elastic Net: C_{123}	0.395	0.151	0.98
Random Forest: C_{123}	0.393	0.152	0.98
GLM: C_{123}	0.392	0.152	0.98
SVM: C_{123}	0.369	0.158	0.94
Random Forest: C_{12}	0.300	0.175	0.85
GAM: C_{12}	0.299	0.175	0.85
Ridge: C_{12}	0.298	0.175	0.85
Lasso: C_{12}	0.298	0.175	0.85
Elastic Net: C_{12}	0.298	0.175	0.85
GLM: C_{12}	0.297	0.176	0.85
SVM: C_{12}	0.259	0.185	0.80
Neural Net: C_1	0.220	0.195	0.76
GLM: C_1	0.219	0.195	0.76
GAM: C_1	0.219	0.195	0.76
Ridge: C_1	0.219	0.195	0.76
Elastic Net: C_1	0.219	0.195	0.76
Lasso: C_1	0.219	0.195	0.76
SVM: C_1	0.082	0.229	0.65
Neural Net: C_{12}	0.000	0.250	0.59
Neural Net: C_{123}	0.000	0.250	0.59
Random Forest: C_1	-0.035	0.259	0.57
Clinical Tree	-1.006	0.501	0.30

When we apply the classification rules to the (cross-validated and final) predicted probabilities from the super learner function, all rules outperform the clinical tree algorithm in terms of improved sensitivity, accuracy, and AUC (Appendix B, Table 5). The 50% and median rules perform very similarly because the median predicted probability in our sample is 49%. The median rule performs the best with respect to cross-validated AUC and was thereby the selected rule. Table 3 shows the substantial classification improvement of the super learner with median rule over the clinical tree algorithm for the full sample. Figure 2 displays observed early stage by decile of predicted probability, showing the final super learner predicted probabilities classified according to the median rule has *perfect prediction* in the lower 30^{th} and upper 30^{th} of the probability distribution.

	Super Learner	Clinical Tree
True Positives	**6761**	**3865**
False Negatives	**490**	**3386**
True Negatives	6881	6678
False Positives	611	814
Sensitivity	**93**	**53**
Specificity	92	89
Accuracy	**93**	**72**

Table 3: Full Sample Classification Results

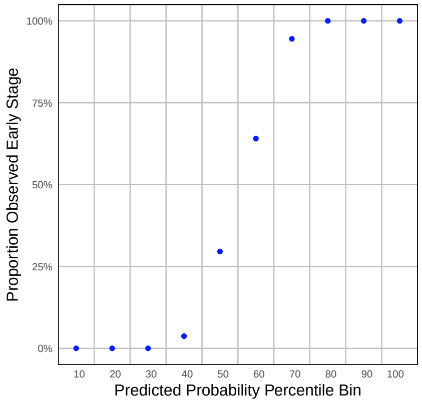

Figure 2: Stage by Predicted Probability

We plot cross-validated AUC by our variable sets C_1, C_{12}, and C_{123} for our algorithms using the median classification rule, including the super learner in each plot for reference, in Figure 3.

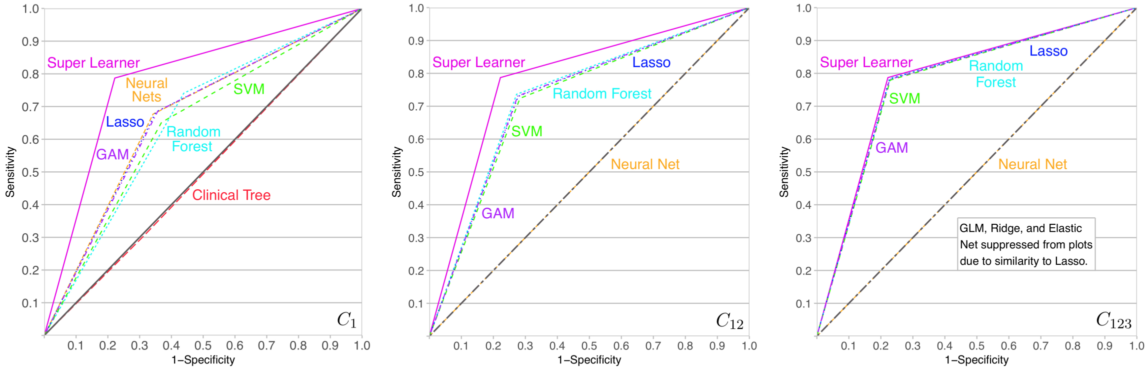

Figure 3: Cross-Validated AUC Plots by Variable Sets $C_{(\cdot)}$

Overall, the largest gains in algorithm performance are driven by the inclusion of additional claims variables, although even using only variable set C_1 (containing only 7 clinical variables) improved on the clinical tree in each of the eight individual algorithm for all metrics, some by at least 2-fold. The addition of the potentially unreliable ICD9 codes for lung cancer type and secondary malignancies (C_3) in variable set C_{123} provided nontrivial improvements over variable set C_{12}. Examining the logistic and lasso regressions using C_{123} provides some insight into the variables driving the stark improvement in sensitivity. The largest coefficients from the logistic regression belong to variables based on the ICD9 malignancy codes, including an indicator for having any malignancy code in the secondary malignancies series: 196 (lymph node sites), 197 (respiratory/digestive sites), or 198 (other sites) or the 199.0 code (malignant neoplasm without specification of site). The lasso selected 75 variables, including those related to resection, radiation, SEER registry region, chemotherapy agents, race, comorbidities, all ICD9 codes except 162.8 (lung cancer at other site), and sex. The largest coefficients from the lasso belong to variables indicating receipt of lobectomy, pneumonectomy, any stereotactic cranial radiation within three months of initial lung cancer chemotherapy, and any lung resection surgery in the three months prior to initial lung cancer chemotherapy. Additionally, some of the best performing candidate algorithms have not been previously used in the literature to stage lung cancer using administrative claims data.

5. Discussion

The development of an algorithm to classify lung cancer stage is needed to allow researchers to use administrative claims data for studying lung cancer patient quality of care and health outcomes. Prior work has been unable to simultaneously achieve 80% sensitivity and specificity. Earlier work evaluating the performance of the clinical tree algorithm (Brooks et al., 2017) echoed the suboptimal performance we confirm for it here. The overall conclusion has been that it is not possible to rely on claims data to conduct rigorous health outcomes research for lung cancer patients. Using an expanded set of variables and algorithms, we demonstrate that ensemble machine learning methods can be used to classify lung cancer patients receiving chemotherapy with 93% sensitivity, 92% specificity, and overall accuracy of 93%.

While the super learner yielded the best performance in terms of cross-validated MSE, several individual algorithms performed with a high degree of relative efficiency, including regression-based techniques. An alternative approach would be to a priori define an "improvement threshold" by which more complex algorithms must outperform in order to be selected. That said, super learner classified 1,034 additional people as true positives and 1,034 additional people as true negatives, resulting in a 14 percentage point improvement in accuracy, compared to the next best performing algorithm based on cross-validated MSE (GAM with C_{123}). An improvement of this level is likely to be clinically meaningful, although may depend on the context and setting the tool will be used.

Future directions for this work include exploring multi-level classification (i.e., stage I/II, stage III, and stage IV) and reducing the number and complexity of features required for accurate performance. This latter advance would maximize the practical utility of the algorithm for health services researchers using claims data to study quality of care and health outcomes. We will also build classification algorithms for other cancer types. These tools will be published on the project website `cancerclas.org` as they are developed and validated.

Acknowledgements

The authors thank Joyce Lii and Robert Wolf for data management and programming assistance as well as Lauren Riedel for research support.

The analyses upon which this publication is based were performed under Contract Number 46244, entitled, “Evaluation of the Oncology Care Model,” sponsored by the Centers for Medicare & Medicaid Services, Department of Health and Human Services. This study used the linked SEER-Medicare database. The interpretation and reporting of these data are the sole responsibility of the authors. The authors acknowledge the efforts of 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 in the creation of the SEER-Medicare database. The collection of cancer incidence data used in this study was supported by the California Department of Public Health as part of the statewide cancer reporting program mandated by California Health and Safety Code Section 103885; the National Cancer Institute’s Surveillance, Epidemiology and End Results Program under contract HHSN261201000140C awarded to the Cancer Prevention Institute of California, contract HHSN261201000035C awarded to the University of Southern California, and contract HHSN261201000034C awarded to the Public Health Institute; and the Centers for Disease Control and Prevention’s National Program of Cancer Registries, under agreement # U58DP003862-01 awarded to the California Department of Public Health. The ideas and opinions expressed herein are those of the author(s) and endorsement by the 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. The authors acknowledge the efforts of 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 in the creation of the SEER-Medicare database.

Research reported in this publication was also supported by the National Institute Of Mental Health of the National Institutes of Health under Award Number T32MH019733. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health

References

G.A. Brooks, M.B. Landrum, and N.L. Keating. Inferring cancer stage from administrative data, March 2017. Report submitted to the Centers for Medicare and Medicaid Innovation.

N. Chawla, K. R. Yabroff, A. Mariotto, T. S. McNeel, D. Schrag, and J. L. Warren. Limited validity of diagnosis codes in Medicare claims for identifying cancer metastases and inferring stage. *Ann Epidemiol*, 24(9):666–672, 2014.

G. S. Cooper, Z. Yuan, K. C. Stange, S. B. Amini, L. K. Dennis, and A. A. Rimm. The utility of Medicare claims data for measuring cancer stage. *Med Care*, 37(7):706–711, 1999.

P. Guan, D. Huang, M. He, and B. Zhou. Lung cancer gene expression database analysis incorporating prior knowledge with support vector machine-based classification method. *Journal of Experimental & Clinical Cancer Research*, 28(1):103, 2009.

M. J. Hassett, D. P. Ritzwoller, N. Taback, N. Carroll, A. M. Cronin, G. V. Ting, D. Schrag, J. L. Warren, M. C. Hornbrook, and J. C. Weeks. Validating billing/encounter codes as indicators of lung, colorectal, breast, and prostate cancer recurrence using 2 large contemporary cohorts. *Med Care*, 52(10):65–73, 2014.

F. Hosseinzadeh, M. Ebrahimi, B. Goliaei, and N. Shamabadi. Classification of lung cancer tumors based on structural and physicochemical properties of proteins by bioinformatics models. *PLoS One*, 7, 2012.

N. Howlader, A.M. Noone, M. Krapcho, D. Miller, K. Bishop, S.F. Altekruse, C.L. Kosary, M. Yu, J. Ruhl, Z. Tatalovich, A. Mariotto, D.R. Lewis, H.S. Chen, E.J. Feuer, and K.A. Cronin (eds.). SEER cancer statistics review 1975-2013. Report, National Cancer Institute, 2016.

K. Konstantina, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, and D.I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8 – 17, 2015. ISSN 2001-0370. doi: http://dx.doi.org/10.1016/j.csbj.2014.11.005. URL http://www.sciencedirect.com/science/article/pii/S2001037014000464.

V.A. Moyer. Screening for lung cancer: U.S. Preventive Services Task Force recommendation statement. *Annals of Internal Medicine*, 160(5):330–338, 2014. doi: 10.7326/M13-2771. URL +http://dx.doi.org/10.7326/M13-2771.

NCCN. National Comprehensive Cancer Network Clinical practice guidelines in oncology. Non-small cell lung cancer. Version 4.2017. https://www.nccn.org/professionals/physician_gls/pdf/nscl.pdf, 2017. Accessed: March 01, 2017.

B. L. Nordstrom, J. L. Whyte, M. Stolar, C. Mercaldi, and J. D. Kallich. Identification of metastatic cancer in claims data. *Pharmacoepidemiol Drug Saf*, 21 Suppl 2:21–28, 2012.

B.L. Nordstrom, J.C. Simeone, K.G. Malley, K.H. Fraeman, Z. Klippel, M. Durst, J.H. Page, and H. Xu. Validation of claims algorithms for progression to metastatic cancer in patients with breast, non-small cell lung, and colorectal cancer. *Frontiers in Oncology*, 6:18, 2016. ISSN 2234-943X. doi: 10.3389/fonc.2016.00018. URL http://journal.frontiersin.org/article/10.3389/fonc.2016.00018.

E.C. Polley, E. LeDell, C. Kennedy, S. Lendle, and M. van der Laan. Superlearner: Super Learner prediction. version 2.0-21. https://cran.r-project.org/web/packages/SuperLearner/, 2016. Accessed: March 24, 2017.

A. L. Potosky, G. F. Riley, J. D. Lubitz, R. M. Mentnech, and L. G. Kessler. Potential for cancer related health services research using a linked Medicare-tumor registry database. *Med Care*, 31(8):732–748, 1993.

R.L. Siegel, K.D. Miller, and A. Jemal. Cancer statistics, 2017. *CA: A Cancer Journal for Clinicians*, 67(1):7–30, 2017. ISSN 1542-4863. doi: 10.3322/caac.21387. URL `http://dx.doi.org/10.3322/caac.21387`.

M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Stat Appl Genet Mol Biol*, 6:Article25, 2007.

H. Wang, Z. Zhou, Y. Li, Z. Chen, P. Lu, W. Wang, W. Liu, and L. Yu. Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18f-fdg pet/ct images. *EJNMMI Research*, 7(1):11, 2017. ISSN 2191-219X. doi: 10.1186/s13550-017-0260-9. URL `http://dx.doi.org/10.1186/s13550-017-0260-9`.

J.L. Warren and K.R. Yabroff. Challenges and opportunities in measuring cancer recurrence in the United States. *JNCI: Journal of the National Cancer Institute*, 107(8):djv134, 2015. doi: 10.1093/jnci/djv134. URL `+http://dx.doi.org/10.1093/jnci/djv134`.

J.L. Whyte, N.M. Engel-Nitz, A. Teitelbaum, G. Gomez Rey, and J.D. Kallich. An evaluation of algorithms for identifying metastatic breast, lung, or colorectal cancer in administrative claims data. *Medical Care*, 53:49–57, 2015. ISSN 1537-1948. doi: 10.1097/MLR.0b013e318289c3fb.

N. Yamagata, Y. Shyr, K.i Yanagisawa, M. Edgerton, T.P. Dang, A. Gonzalez, S. Nadaf, P. Larsen, J.R. Roberts, J.C. Nesbitt, R. Jensen, S. Levy, J.H. Moore, J.D. Minna, and D.P. Carbone. A training-testing approach to the molecular classification of resected non-small cell lung cancer. *Clinical Cancer Research*, 9(13):4695–4704, 2003.

K. H. Yu, C. Zhang, G. J. Berry, R. B. Altman, C. Re, D. L. Rubin, and M. Snyder. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat Commun*, 7:12474, 2016.

Appendix A

Table 4: SEER-Medicare Data Features

Variable(s)	Notes
Age (years)	integer, 28–101
Sex	binary
Census Tract Median Income	integer, $7,104–200,000
Census Tract Non-HS Grad	continuous, 0–79.91%
Census Tract Below Poverty	continuous, 0–82.05%
Race	indicators *black*, *asian*, *hispanic*, *other*; reference *white*
SEER Registry Location	16 indicators
ICD9 162.0 Rate	continuous; % lung cancer codes trachea site
ICD9 162.2 Rate	continuous; % lung cancer codes main bronchus site
ICD9 162.3 Rate	continuous; % lung cancer codes upper lobe site
ICD9 162.4 Rate	continuous; % lung cancer codes middle lobe site
ICD9 162.5 Rate	continuous; % lung cancer codes lower lobe site
ICD9 162.8 Rate	continuous; % lung cancer codes other site
ICD9 162.9 Rate	continuous; % lung cancer codes unspecified site
ICD9 Secondary Malignancy	binary; any 196.xx, 197.xx, 198.xx, and 199.0 codes
ICD9 Secondary Malignancy Total	integer, 0–93; total of 196.xx, 197.xx, 198.xx, 199.0 codes
ICD9 196.xx Rate	continuous; % metastases codes lymph node sites
ICD 197.xx Rate	continuous; % metastases codes respiratory/digestive sites
ICD9 198.xx Rate	continuous; % metastases codes other sites
ICD9 199.0 Rate	continuous; % metastases codes unspecified site
Claim Types	5 binary[*]
Claim Type Counts	5 integers[*]; total for each claim type
Lung Resection Surgery	4 binary[**]
Diagnosis Code at 1st Radiation	indicators *lung cancer*, *secondary malignancy*, *other*; reference *no non-stereotactic radiation*
Radiation Types	2 binary[†]
Radiation Code Totals	2 integers[†]; total for each radiation type
Radiation Fractions Type	4 binary[‡]
Radiation Fractions Totals	3 integers[‡]; total for first 3 fractions types
Radiation Before Surgery	indicators *1st before surgery*, *1st after surgery*, *no surgery*; reference *none within 3 months of chemo*
Chemotherapy Type	17 binary[§]
Chemotherapy Drug Totals	13 integers; # treatment days for each chemotherapy drug[§]
Comorbidities	14 binary[¶]

[*] # outpatient evaluation and management (E&M) claims; inpatient E&M claims; critical care claims; hospital discharges; chemotherapy treatment dates. [**] Any resection surgery; lobectomy; penumonectomy; segmental. [†] Non-brain stereotactic radiation (77373, 77435); brain stereotactic radiation (77371, 77372, 77432). [‡] Radiation fractions within 3 mo of 1st chemo; non-stereotactic fractions within 180 d of 1st chemo; non-stereotactic fractions within 60 d of 1st radiation; ≥ 20 fractions starting ≤ 7 days before 1st lung cancer chemo. [§] Cisplatin, carboplatin, paclitaxel, docetaxel, pemetrexed, gemcitabine, vinorelbine, bevacizumab, etoposide, irinotecan, topotecan, trastuzumab, & unclassified drug. Add'l chemo indicators include: no receipt of lung cancer chemo; advanced NSCLC chemo; small-cell chemo agents & platinums only; targeted agents. [¶] Dementia; acute myocardial infarction; ischemic heart disease; stroke/TIA; atrial fibrillation; hip/pelvic fracture; heart failure; hypertension; hyperlipidemia; diabetes; asthma; COPD; depression; chronic kidney disease.

Appendix B

Table 5: Super Learner and Clinical Tree Classification Performance

	Fixed			Percentile					**Clinical**
	40	**50**	60	30^{th}	40^{th}	**50**th	60^{th}	70^{th}	**Tree**
Full Sample									
Sensitivity	97	**93**	85	100	99	**93**	80	61	**53**
Specificity	85	**92**	97	59	78	**92**	99	100	**89**
Accuracy	91	**92**	91	79	88	**93**	90	81	**72**
AUC	91	**92**	91	80	89	**93**	90	80	**71**
Cross-Validated									
Sensitivity	87	**78**	67	95	88	**79**	67	53	**31**
Specificity	67	**78**	87	54	67	**78**	87	93	**69**
Accuracy	77	**78**	77	74	77	**78**	77	73	**50**
AUC	77	**78**	77	74	77	**78**	77	73	**50**