original report

Classifying Stage IV Lung Cancer From Health Care Claims: A Comparison of Multiple Analytic Approaches

Gabriel A. Brooks, MD, MPH[1]; Savannah L. Bergquist, MSc[2]; Mary Beth Landrum, PhD[3]; Sherri Rose, PhD[3]; and Nancy L. Keating, MD, MPH[3]

abstract

PURPOSE Cancer stage is a key determinant of outcomes; however, stage is not available in claims-based data sources used for real-world evaluations. We compare multiple methods for classifying lung cancer stage from claims data.

METHODS Our study used the linked SEER-Medicare data. The patient samples included fee-for-service Medicare beneficiaries diagnosed with lung cancer from 2010 to 2011 (development cohort) and 2012 to 2013 (validation cohort) who received chemotherapy. Classification algorithms considered Medicare Part A and B claims for care in the 3 months before and after chemotherapy initiation. We developed a clinical algorithm to predict stage IV (*v* I to III) cancer on the basis of treatment patterns (surgery, radiotherapy, chemotherapy). We also considered an ensemble of claims-based machine learning algorithms. Classification methods were trained in the development cohort, and performance was measured in both cohorts. The SEER data were the gold standard for cancer stage.

RESULTS Development and validation cohorts included 14,760 and 14,620 patients with lung cancer, respectively. Validation analyses assessed clinical, random forest, and simple logistic regression algorithms. The best performing classifier within the development cohort was the random forests, but this performance was not replicated in validation analysis. Logistic regression had stable performance across cohorts. Compared with the clinical algorithm, the 14-variable logistic regression algorithm demonstrated higher accuracy in both the development (77% *v* 71%) and validation cohorts (77% *v* 73%), with improved specificity for stage IV disease.

CONCLUSION Machine learning algorithms have potential to improve lung cancer stage classification but may be prone to overfitting. Use of ensembles, cross-validation, and external validation can aid generalizability. Degradation of accuracy between development and validation cohorts suggests the need for caution in implementing machine learning in research or care delivery.

INTRODUCTION

Cancer stage is a critical determinant of health outcomes and spending for patients with cancer, as well as a key criterion for determining appropriate cancer treatments. As health care providers and policy makers increasingly seek to use existing health care data to gain insights into health care quality and costs, the unavailability of cancer staging information in administrative data is a substantial obstacle. Without cancer stage information, claims-based analyses of cancer outcomes lack a critical variable that mediates both cancer treatments and health outcomes.

To improve the utility of administrative data sources, prior studies have proposed various approaches to infer cancer stage from information readily available in claims data, including diagnosis and procedure codes.[1-4] These approaches rely heavily on secondary site diagnosis codes, which indicate the presence of a distant metastatic site (eg, International Classification of Diseases, Ninth Revision, Clinical Modification [ICD-9-CM] code 197.7: malignant neoplasm of liver, secondary). Because use of these secondary site codes is not required for payment (and is therefore variable), these approaches have shown good specificity (generally ≥ 80%) but poor sensitivity (≤ 60%) for identifying patients with advanced cancer. Correspondingly poor accuracy has hampered the uptake of stage inference algorithms. Accordingly, analyses of cancer care quality, outcomes, and costs have relied heavily on linkages of administrative data with clinical data from cancer registries, such as the linked SEER-Medicare data.[5] The SEER-Medicare data are limited by their focus on older, fee-for-service Medicare beneficiaries living in SEER regions and by the lag time required to produce the data linkage. Nevertheless,

ASSOCIATED CONTENT

Appendix

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on March 4, 2019 and published at ascopubs.org/journal/cci on May 9, 2019: DOI https://doi.org/10.1200/CCI.18.00156

CONTEXT

Key Objective

To examine multiple analytic approaches for classifying lung cancer stage group (stage IV *v* I to III) using health care claims among patients receiving chemotherapy for lung cancer.

Knowledge Generated

We compared the accuracy of an algorithm on the basis of clinical logic with multiple machine learning algorithms for classifying lung cancer stage group from Medicare claims records. The selected classifiers (random forest and logistic regression) demonstrated improved accuracy compared with the clinical algorithm, and logistic regression exhibited the greatest stability (77% accuracy in both development and validation analyses).

Relevance

The stage classification approach described here can be used in stratified or risk-adjusted analyses of real-world health care delivery and health outcomes among patients receiving chemotherapy for lung cancer. Our results are most relevant to analyses of Medicare claims records; additional study is needed to test generalizability in other populations (eg, commercially insured populations).

these data have been tremendously fruitful for understanding real-world cancer care delivery and outcomes.

In this report, we describe several approaches for classifying cancer stage group from health care claims data among fee-for-service Medicare beneficiaries. We focus on patients receiving chemotherapy within 6 months of cancer diagnosis, because this population has particular policy relevance in the context of episode-based payment models that are structured around chemotherapy receipt.[6,7] The first approach involves specification of a clinical algorithm that assigns cancer stage on the basis of procedure and associated diagnosis codes for chemotherapy, radiotherapy, and lung cancer resection surgeries. In the latter approaches, we apply machine learning techniques for stage classification, using a curated and clinically informed data set of demographic, diagnostic, and treatment-related variables derived from Medicare claims.

METHODS

Data Source

We used SEER-Medicare linked data for all analyses. The SEER program of the National Cancer Institute collects uniformly reported data from population-based cancer registries, including cancer site, stage, month of diagnosis, and other clinical variables from areas covering 28% of the United States.[5] Since 1991, the National Cancer Institute has linked SEER data with Medicare administrative data for more than 94% of SEER registry patients diagnosed with cancer at age 65 years or older.[8] Medicare data used in this analysis included fee-for-service inpatient, outpatient, provider carrier (Physician/Supplier Part B), and durable medical equipment claims.

Study Sample

We identified all fee-for-service Medicare beneficiaries in the SEER-Medicare linked data with a new diagnosis of lung cancer from 2010 to 2013 (the most recent 4-year period for which data were available at the time this study was conducted). We then restricted our study sample to patients who had an index chemotherapy claim within 6 months of cancer diagnosis that was associated with an ICD-9-CM diagnosis code for lung cancer. We identified qualifying chemotherapy agents on the basis of the chemotherapy trigger list developed for the Oncology Care Model (OCM), a Centers for Medicare and Medicaid Services (CMS) payment model designed around 6-month chemotherapy treatment episodes.[9] Patients were excluded from the analytic sample if they had incomplete cancer stage information or if they were not enrolled in fee-for-service Medicare for the entire treatment ascertainment period (3 months before through 3 months after the date of first chemotherapy within 6 months of diagnosis). Cohort selection is described further in Appendix Figure A1.

SEER-derived staging variables on the basis of American Joint Committee on Cancer (version 6) were used as the gold standard for assigning lung cancer stage group.[10] We used SEER collaborative stage fields (local, regional, metastatic) to assign stage for 1.4% of patient cases with missing data for the American Joint Committee on Cancer staging variables.

The study sample was split into two cohorts for algorithm development (2010 to 2011 diagnoses) and validation (2012 to 2013 diagnoses). Using successive time periods for algorithm development and validation approximates how these or similar classification algorithms would be implemented for common real-world uses.

Algorithm Development

Clinical algorithm. The clinical algorithm classified patients with lung cancer into cancer stage groups on the basis of treatments received in the 3 months before and after the chemotherapy trigger date (classification period). Specifically, we examined receipt and timing of treatment with chemotherapy (including specific agents), radiation

therapy (including number of fractions and use of cranial irradiation), and surgery (pneumonectomy, lobectomy, or wedge resection). Receipt of medical treatment was determined from Medicare claims files, using ICD-9-CM procedure codes, Healthcare Common Procedure Coding System procedure codes, and drug codes (Healthcare Common Procedure Coding System J codes for medications administered in outpatient and office-based settings). The clinical algorithm was iteratively revised within a 50% subset of the development cohort until the research team determined that further optimization was impractical, as measured by joint sensitivity and specificity. These analyses were performed in SAS software (version 9.4; SAS Institute, Cary, NC).

The final clinical algorithm classified patients into six terminal branches, with each branch representing a treatment approach for either stage I to III or stage IV lung cancer (Fig 1). For example, patients undergoing lung cancer resection surgery in the 3 months before or after starting chemotherapy were classified as having stage I to III lung cancer, as were patients receiving 20 or more fractions of concurrent chemoradiotherapy. Patients who were treated with chemotherapy only, without surgery or extended-fraction radiotherapy, were classified as having stage 4 disease. Specifications for the clinical algorithm are listed in Appendix Table A1.

Machine learning algorithms. We deployed multiple machine learning algorithms to identify a high-performing classifier for lung cancer stage group.[11,12] The ensemble of algorithms considered logistic regressions, random forests, generalized additive regressions, classification trees, and pruned classification trees, using 10-fold cross-validation. Because of our interest in practical use and simplicity, our ensemble aimed to select the single best individual algorithm rather than a complex weighted average of algorithms. The development cohort input data set for the machine learning ensemble contained 102 variables derived from or linkable with the Medicare claims data, using the same measurement period as the clinical algorithm (3 months before and after the initial chemotherapy date). Variables included each of the classification nodes of the clinical algorithm, as well as additional variables for demographic characteristics (age, sex, race/ethnicity, geographic region, and census tract–level variables characterizing median household income, proportion of residents without a high school education, and proportion of residents living in poverty), lung cancer–related diagnosis codes (eg, number of 162.x diagnosis codes), secondary malignancy codes (196.x to 198.x, 199.0), evaluation and management codes, and indicators for receipt of specific chemotherapy agents, radiation therapy, and lung cancer–related surgeries. We used the CMS chronic conditions warehouse to flag comorbid conditions before the first chemotherapy date. Variables in the machine learning data set are listed in Appendix Tables A2 and A3.

To generate relatively parsimonious algorithms with greater potential for practical use, we implemented a variable reduction approach using the least absolute shrinkage and selection operator (LASSO) to select variable sets within each cross-validation fold.[13,14] We investigated six different thresholds for the maximum number of variables selected by the LASSO, including at most 10, 15, 20, 30, 40, or 50 variables. These reduced subsets of variables were then provided to each of the included algorithms, some of which might perform additional variable selection on the subset. Thus, each of the candidate machine learning algorithms was included six times in the ensemble, once with each variable threshold. A previous ensemble approach for lung cancer stage classification did not consider parsimony and produced a complex weighted average of five algorithms, relying on more than 100 variables.[15] These analyses were performed in the R statistical programming language (version 3.1.0; R Foundation, Vienna, Austria).

Analysis

Candidate algorithms from the machine learning ensemble were evaluated based on cross-validated area under the curve (AUC), sensitivity, specificity, and accuracy for classifying stage IV versus stage I to III disease. Results from the development cohort for sensitivity, specificity, and accuracy were calculated for all data for our classification approaches, rather than their respective internal holdout data, as well as for the validation cohort. This yielded the most effective comparison with the clinical literature, because these evaluation metrics are typically reported for an entire study cohort. We present two classifiers from among the machine learning approaches for comparison with the clinical algorithm, balancing parsimony and performance of the fixed algorithms in the two cohorts.

RESULTS

Population

There were 14,760 patients with a new lung cancer diagnosis from 2010 to 2011 (development cohort) and 14,620 patients from 2012 to 2013 (validation cohort). The mean ages in the development and validation cohorts were 72.1 and 71.9 years, respectively. Additional demographic information is listed in Table 1 and Appendix Tables A2 and A3. The proportion of patients with stage 4 lung cancer (based on SEER registry data) was 50.8% in the development cohort and 52.0% in the validation cohort.

Clinical Algorithm

In the development cohort, the clinical algorithm exhibited an overall accuracy of 71% and a specificity of 53% for classification of stage 4 cancer (Table 2). The performance of the clinical algorithm was similar in the validation cohort (accuracy, 73%; specificity, 55%).

Machine Learning Algorithms

On the basis of our variable reduction strategy, the LASSO variable selection tool identified six candidate variable sets

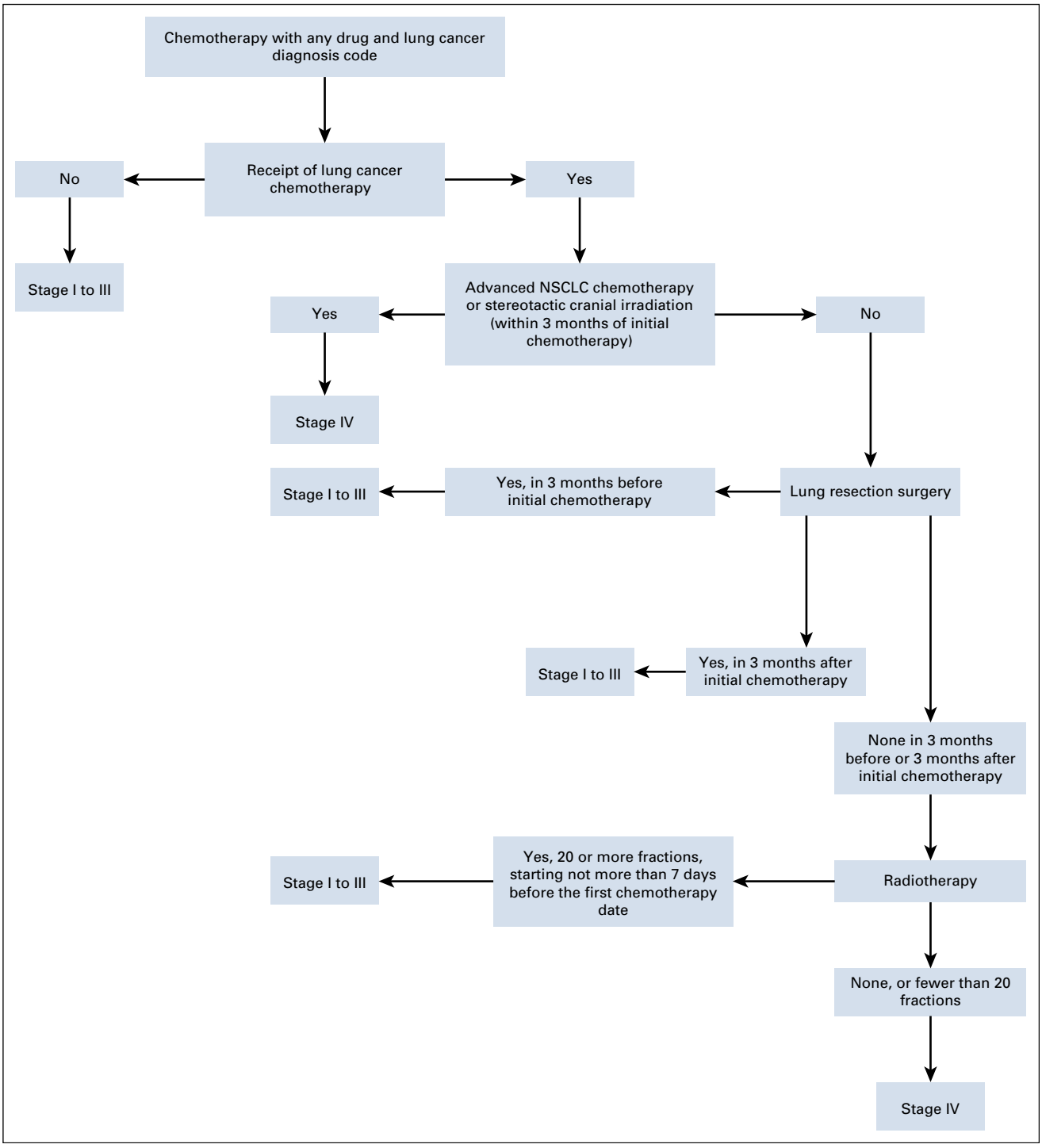

FIG 1. Schematic of clinical algorithm. NSCLC, non–small-cell lung cancer.

of eight, 12, 14, 23, 33, and 44 variables. The random forests were the best performing classifiers for stage group at all covariate thresholds in the development cohort. All six random forests had similar performance with respect to cross-validated AUC (77% to 78%), as did other algorithms (eg, logistic regression cross-validated AUC was 76% to 77%). Only the algorithms with 33 and 44 variables achieved an accuracy of 90% or greater in the development cohort. The 14-variable random forest algorithm exhibited an overall accuracy of 81%, with specificity of 84% in the development cohort. However, the performance of this algorithm deteriorated in the validation cohort, with an accuracy of 78% (Table 2). In contrast, the logistic regressions had stable performance across all covariate

TABLE 1. Demographic Characteristics of the Study Population

Characteristic	Development Cohort (2010 to 2011)	Validation Cohort (2012 to 2013)
Total No.	14,760	14,620
Mean (SD) age, years	72.1 (7.7)	71.9 (7.7)
Sex, %		
Male	54.6	53.2
Female	45.4	46.8
Race/ethnicity, %		
White	82.7	81.4
Black	8.8	8.9
Hispanic	4.2	4.4
Asian	3.9	4.8
Other	0.3	0.5
Average (SD) median income of zip code of residence, $	60,619 (28,580)	60,766 (28,933)
Mean (SD) not graduating high school, %	20.2 (12.9)	14.8 (10.8)
Mean (SD) residents living below poverty level, %	11.7 (9.5)	14.3 (10.8)
Region, %		
Northeast	20.3	20.2
Midwest	13.6	13.2
West	34.2	34.5
South	31.9	32.2

Abbreviation: SD, standard deviation.

thresholds and across the development and validation cohorts, while also improving on the clinical algorithm with respect to specificity (78% *v* 55%) and accuracy (77% *v* 73%).

The covariates from the 14-variable algorithms are listed in Table 3 and include the count of all secondary malignancy codes, ratio measures of lung cancer and secondary malignancy diagnosis codes, indicators of specific patterns of radiotherapy, surgery, and chemotherapy receipt, and an indicator for residence in the US Midwest. In sensitivity analyses, we investigated whether other algorithms with more or fewer variables fit in the development cohort yielded improved performance in the validation cohort. No other variable thresholds or algorithms improved meaningfully on the findings for the 14-variable logistic regression (results not shown).

Comparison of Clinical and Machine Learning Approaches

Given our goals of accuracy and parsimony, we selected the logistic regression with 14 variables as the best algorithm for overall performance (inclusive of accuracy, sensitivity, and specificity). However, the logistic regression had lower specificity than the clinical algorithm in both cohorts. The three-way agreement of the clinical algorithm, logistic regression, and SEER-recorded stage in the validation cohort is summarized in Table 4.

DISCUSSION

We demonstrated that claims-based algorithms can classify lung cancer stage group (stage IV *v* I to III) with good sensitivity, specificity, and accuracy among patients receiving initial chemotherapy. The machine learning algorithms modestly outperformed the clinical classification algorithm in both development and validation cohorts. Performance of the random forests algorithm declined nontrivially in the validation cohort (compared with performance in the development cohort), whereas a simple logistic regression showed stability across development and validation cohorts. Secondary site diagnosis codes figured prominently in both the random forests and logistic regression algorithms, despite moderate to poor sensitivity for detection of advanced-stage disease in prior studies.[3,4]

Both our clinical algorithm and machine learning approaches were designed to incorporate oncology knowledge. Notably, we imposed considerable structure onto the machine learning development in the preanalytic phase, categorizing codes for related clinical concepts, creating variables for counts of procedure and diagnosis codes, and defining ratio measures (eg, percentage of lung cancer diagnosis codes that were malignant neoplasm of upper lobe, bronchus, or lung). As such, our machine learning approaches represent a fusion of applied clinical information with data-adaptive statistical learning. This strategy for building computational tools is likely more efficient than an unstructured approach that ignores clinical input and may be more robust to changes in billing, coding, and practice patterns.

Our findings help inform the use of machine learning classification algorithms to enhance claims-based analyses of cancer outcomes. By extension, our approach can support the concept of a "learning health care information system for cancer,"[16(p235)] with the ultimate goal of facilitating knowledge generation from observational data. Additional potential roles for a high-fidelity claims-based stage classification algorithm lie in the domains of quality measurement and risk adjustment; both are likely to be critical components of value-based payment approaches for cancer care.

This work was motivated in substantial part by the OCM, an episode-based payment model run by CMS.[6,9] The OCM is a payment model built on the scaffolding of fee-for-service medicine, with the potential for performance-based payments to oncology practices that meet quality standards and reduce total Medicare spending below the target price for a 6-month episode. To evaluate whether the incentive structures of the model affect quality and outcomes of care, it will be necessary to conduct stage-adjusted analyses of treatment patterns and patient outcomes. It is our intent that machine learning approaches for classifying cancer

TABLE 2. Comparative Performance Clinical and Machine-Learning Classification Algorithms

Algorithm	Sensitivity % (95% CI)*	Specificity % (95% CI)*	Accuracy % (95% CI)
Development cohort			
Clinical algorithm	89 (88 to 90)	53 (52 to 54)	71 (71 to 72)
Random forest algorithms, No. of variables			
8	71 (71 to 72)	84 (84 to 85)	77 (77 to 78)
12	74 (74 to 75)	84 (84 to 85)	79 (79 to 80)
14	79 (79 to 80)	84 (84 to 85)	81 (81 to 82)
23	84 (84 to 85)	89 (89 to 90)	86 (86 to 87)
33	92 (92 to 93)	96 (96 to 97)	94 (94 to 95)
44	99 (99 to 99)	99 (99 to 99)	99 (99 to 99)
Logistic regressions, No. of variables			
8	69 (69 to 70)	84 (84 to 85)	76 (76 to 77)
12	72 (72 to 73)	81 (81 to 82)	77 (77 to 78)
14	76 (76 to 77)	77 (77 to 78)	77 (77 to 78)
23	77 (77 to 78)	78 (78 to 79)	78 (78 to 79)
33	77 (77 to 78)	78 (78 to 79)	78 (78 to 79)
44	77 (77 to 78)	78 (78 to 79)	78 (78 to 79)
Validation cohort			
Clinical algorithm	90 (89 to 91)	55 (54 to 56)	73 (72 to 74)
Random forest algorithm†	76 (75 to 77)	80 (79 to 81)	78 (77 to 79)
Logistic regression†	77 (76 to 78)	78 (77 to 79)	77 (76 to 78)

*Sensitivity and specificity are reported in reference to stage IV lung cancer.
†Validation analyses used the 14-variable machine-learning algorithms.

stage can be used in the evaluation of the OCM and other value-based payment programs, helping to evaluate the quality and outcomes of care delivered within these models. In this context, it is relevant to ask if our classification algorithm is good enough to be deployed. The logistic regression successfully classified stage group in 77% of patients in the validation cohort, with sensitivity and specificity both exceeding 75%, comparing favorably with previously reported stage classification approaches.[1-4] Even so, 23% of patients were misclassified according to the gold standard of SEER staging. We contend that our stage classification tool, although imperfect, nevertheless provides useful, valuable context for understanding and contextualizing the outcomes of patients with lung cancer. We strongly support additional study to confirm (and improve) the robustness, generalizability, and adaptability of our classification tool.

A strength of this analysis is its use of the SEER-Medicare linked data as the gold standard for cancer stage. At present, SEER-Medicare is essentially the only large-scale US data source (and one of few large data sources internationally) where cancer stage information is reliably linked to clinical claims and survival data. The unique features of the SEER-Medicare data have enabled sophisticated analyses of cancer outcomes in well-defined, real-world populations. Because the SEER-Medicare data include large numbers of patients with incident cancers, they permit analyses of rare patient subgroups and outcomes. By using SEER-Medicare data to develop algorithms for stage classification, we further extend the potential value of SEER-Medicare data to help classify cancer stage in other claims-based data sources, such as the larger, unlinked Medicare data and claims-based data sets derived from commercial insurers.

Our analysis has several limitations. Medicare administrative claims are generated for billing purposes and may contain incomplete, unverified, or incorrect diagnostic information. The machine learning algorithms rely on a broad array of diagnosis and procedure codes, including the secondary malignancy codes (malignant neoplasm of secondary site). When used alone, these secondary malignancy codes have shown poor sensitivity for identification of advanced-stage disease.[1-4] However, secondary malignancy codes represent only one of multiple input sets for our machine learning algorithms. Moreover, our machine learning approach differs substantially from previously described algorithms.

Our analysis was limited to Medicare beneficiaries living in SEER areas receiving initial chemotherapy for lung cancer,

TABLE 3. Selected Variables for the 14-Covariate Machine-Learning Algorithms

Variable Description
Proportion of all lung* and secondary malignancy† diagnosis codes that are in the 198 series (secondary malignant neoplasm of other specified sites)
Count of all secondary malignancy diagnosis codes*
Count of nonstereotactic radiation fractions delivered within 60 days of first radiation treatment
Proportion of all lung cancer diagnosis codes† that are 162.3 (malignant neoplasm of upper lobe, bronchus, or lung)
Count of outpatient evaluation and management claims
Radiation therapy, 20 or more fractions, beginning not more than 7 days before initial lung chemotherapy
Proportion of all lung* and secondary malignancy† diagnosis codes that are in the 197 series (secondary malignant neoplasm of respiratory and digestive systems)
Lung resection surgery in 3 months before initial lung cancer chemotherapy
Lobectomy lung resection type (pneumonectomy, lobectomy, or wedge/segmental)
Small-cell chemotherapy agents and platinum chemotherapies only in 3 months after first lung cancer chemotherapy
Bevacizumab or stereotactic cranial irradiation within 3 months of initial lung cancer chemotherapy
Any cisplatin
Diagnosis code associated with first nonstereotactic radiation delivery code (among patients with radiation therapy), secondary malignancy†
Region of residence Midwest

*International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) diagnosis code of 162.x.

†ICD-9-CM diagnosis codes of 196.x, 197.x, 198.x, and 199.0.

and additional research is needed to assess the performance of these algorithms among commercially insured patients, contemporaneous patient populations, patients receiving treatment for recurrent (rather than incident) disease, and patients who do not receive chemotherapy. The emergence of new monoclonal antibody immunotherapies, including nivolumab, pembrolizumab, atezolizumab, and durvalumab, is of particular salience in lung cancer.[17-20] First approved in 2014, immunotherapy drugs are transforming lung cancer treatment, particularly for stage IV disease. It is uncertain how our classification approach will perform in the context of these treatments. Nevertheless, the machine learning algorithms we report here rely substantially on diagnosis codes and less on chemotherapy treatment codes.

The algorithms may also be sensitive to changes in billing patterns (as reflected in claims data). One recent change is the transition from the ICD-9-CM diagnosis coding system (extant during our evaluation period) to the current standard of ICD-10-CM. Fortunately, the cancer diagnosis codes used in this analysis can be cross-walked to the newer ICD-10 system without substantial ambiguity. Finally, we focused on classifying patients as having stage IV versus stage I to III lung cancer, but treatment patterns and outcomes also differ substantially for patients with stage 1 to 2 versus stage 3 cancers, as well as by lung cancer histology (eg, adenocarcinoma, squamous cell, or small cell). These differences in treatment patterns, such as the use of combined-modality chemoradiotherapy with or without durvalumab as a principal treatment for stage 3 non–small-cell lung cancer, offer the possibility that further, more granular claims-based subclassification may be feasible. However, additional research is needed to assess the ability of machine learning algorithms to further classify patients with lung cancer into more detailed, clinically relevant subgroups.

In conclusion, validated algorithms hold promise to classify cancer stage using claims data without linked registry data. As such, algorithms similar to those reported here could serve a facilitating role in building a learning health care information system, providing necessary structure for clinically relevant, real-world analyses of cancer care delivery processes, quality measures, and clinical outcomes. Ongoing evaluation and updating will be necessary for tools such as these to assess and refit the classification estimators in the context of changing treatment patterns and care delivery settings.

TABLE 4. Three-Way Agreement of the Clinical Algorithm, Logistic Regression, and SEER-Recorded Stage

		No. (%)	
Classified Stage		**SEER Stage (gold standard)**	
Clinical Algorithm	**Logistic Regression**	**Stage I-III**	**Stage IV**
Stage I-III	I-III	3,697 (52.7)	586 (7.7)
Stage I-III	IV	153 (2.2)	185 (2.4)
Stage IV	I-III	1756 (25)	1,161 (15.3)
Stage IV	IV	1,415 (20.2)	5,667 (74.6)

AFFILIATIONS

[1]Geisel School of Medicine, Lebanon, NH
[2]Harvard University, Cambridge, MA
[3]Harvard Medical School, Boston, MA

This study used the linked SEER-Medicare database. The interpretation and reporting of these data are the sole responsibility of the authors. The ideas and opinions expressed herein are those of the authors, 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 it be inferred.

CORRESPONDING AUTHOR

Gabriel A. Brooks, MD, MPH, Dartmouth Hitchcock Medical Center, 1 Medical Center Drive, Lebanon, NH 03756; Twitter: @gabe_a_brooks; e-mail: gabriel.a.brooks@hitchcock.org.

PRIOR PRESENTATION

Presented in part in abstract form at the 2018 Annual Meeting of the American Society of Clinical Oncology, Chicago, IL, June 1-5, 2018.

SUPPORT

Supported in part by Contract No. HHSM-500-2014-00026I from the Centers for Medicare and Medicaid Services, US Department of Health and Human Services, under which analyses on which this report is based were performed; by Grant No. K24CA181510 from the National Cancer Institute (N.L.K.); and by a Career Development Award from the Cancer Conquer Foundation (G.A.B.). 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 SEER program under Contract No. HHSN261201000140C awarded to the Cancer Prevention Institute of California, Contract No. HHSN261201000035C awarded to the University of Southern California, and Contract No. HHSN261201000034C awarded to the Public Health Institute; and the Centers for Disease Control and Prevention National Program of Cancer Registries under Agreement No. U58DP003862-01 awarded to the California Department of Public Health.

AUTHOR CONTRIBUTIONS

Conception and design: Gabriel A. Brooks, Mary Beth Landrum, Sherri Rose, Nancy L. Keating
Financial support: Nancy L. Keating
Administrative support: Nancy L. Keating
Collection and assembly of data: Nancy L. Keating
Data analysis and interpretation: All authors
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/jco/site/ifc.

Gabriel A. Brooks
Consulting or Advisory Role: Abt Associates, CareCentrix

Sherri Rose
Consulting or Advisory Role: Kantar Health
No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

We acknowledge the efforts of the National Cancer Institute; the Office of Research, Development and Information of the Centers for Medicare and Medicaid Services; Information Management Services; and the SEER program tumor registries in the creation of the SEER-Medicare database. We are grateful to Michael Liu, MD, PhD, Robert Wolf, MS, and Joyce Lii, MS, for expert statistical programming assistance; Lauren Riedel for administrative assistance; and Barbara J. McNeil, MD, PhD, and Andrea Hassol, MPH, for helpful comments on an earlier draft of the manuscript.

REFERENCES

1. Cooper GS, Yuan Z, Stange KC, et al: The utility of Medicare claims data for measuring cancer stage. Med Care 37:706-711, 1999
2. Nordstrom BL, Whyte JL, Stolar M, et al: Identification of metastatic cancer in claims data. Pharmacoepidemiol Drug Saf 21:21-28, 2012 (suppl 2)
3. Chawla N, Yabroff KR, Mariotto A, et al: Limited validity of diagnosis codes in Medicare claims for identifying cancer metastases and inferring stage. Ann Epidemiol 24:666-672, 672.e1-2, 2014
4. Whyte JL, Engel-Nitz NM, Teitelbaum A, et al: An evaluation of algorithms for identifying metastatic breast, lung, or colorectal cancer in administrative claims data. Med Care 53:e49-e57, 2015
5. Warren JL, Klabunde CN, Schrag D, et al: Overview of the SEER-Medicare data: Content, research applications, and generalizability to the United States elderly population. Med Care 40:IV-3-IV-18, 2002 (suppl)
6. Kline RM, Bazell C, Smith E, et al: Centers for Medicare and Medicaid Services: Using an episode-based payment model to improve oncology care. J Oncol Pract 11:114-116, 2015
7. Newcomer LN, Gould B, Page RD, et al: Changing physician incentives for affordable, quality cancer care: Results of an episode payment model. J Oncol Pract 10:322-326, 2014
8. Potosky AL, Riley GF, Lubitz JD, et al: Potential for cancer related health services research using a linked Medicare-tumor registry database. Med Care 31:732-748, 1993
9. Centers for Medicare and Medicaid Services: Oncology Care Model. https://innovation.cms.gov/initiatives/oncology-care/
10. Greene FL, Page DL, Fleming ID, et al: AJCC Cancer Staging Manual (ed 6). New York, NY, Springer, 2002
11. van der Laan MJ, Polley EC, Hubbard AE: Super learner. Stat Appl Genet Mol Biol 6:e25, 2007
12. van der Laan MJ, Rose S: Targeted Learning: Causal Inference for Observational and Experimental Data. New York, NY, Springer Science & Business Media, 2011
13. Tibshirani R: Regression shrinkage and selection via the Lasso. J Roy Stat Soc B Met 58:267-288, 1996
14. Rose S, Bergquist SL, Layton TJ: Computational health economics for identification of unprofitable health care enrollees. Biostatistics 18:682-694, 2017

15. Bergquist SL, Brooks GA, Keating NL, et al: Classifying lung cancer severity with ensemble machine learning in health care claims data. Proc Mach Learn Res 68:25-38, 2017
16. Levit L, Balogh E, Nass S, et al: Delivering High-Quality Cancer Care: Charting a New Course for a System in Crisis. Washington, DC, Institute of Medicine, 2013
17. Borghaei H, Paz-Ares L, Horn L, et al: Nivolumab versus docetaxel in advanced nonsquamous non–small-cell lung cancer. N Engl J Med 373:1627-1639, 2015
18. Garon EB, Rizvi NA, Hui R, et al: Pembrolizumab for the treatment of non-small-cell lung cancer. N Engl J Med 372:2018-2028, 2015
19. Antonia SJ, Villegas A, Daniel D, et al: Durvalumab after chemoradiotherapy in stage III non-small-cell lung cancer. N Engl J Med 377:1919-1929, 2017
20. Fehrenbacher L, Spira A, Ballinger M, et al: Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): A multicentre, open-label, phase 2 randomised controlled trial. Lancet 387:1837-1846, 2016

APPENDIX

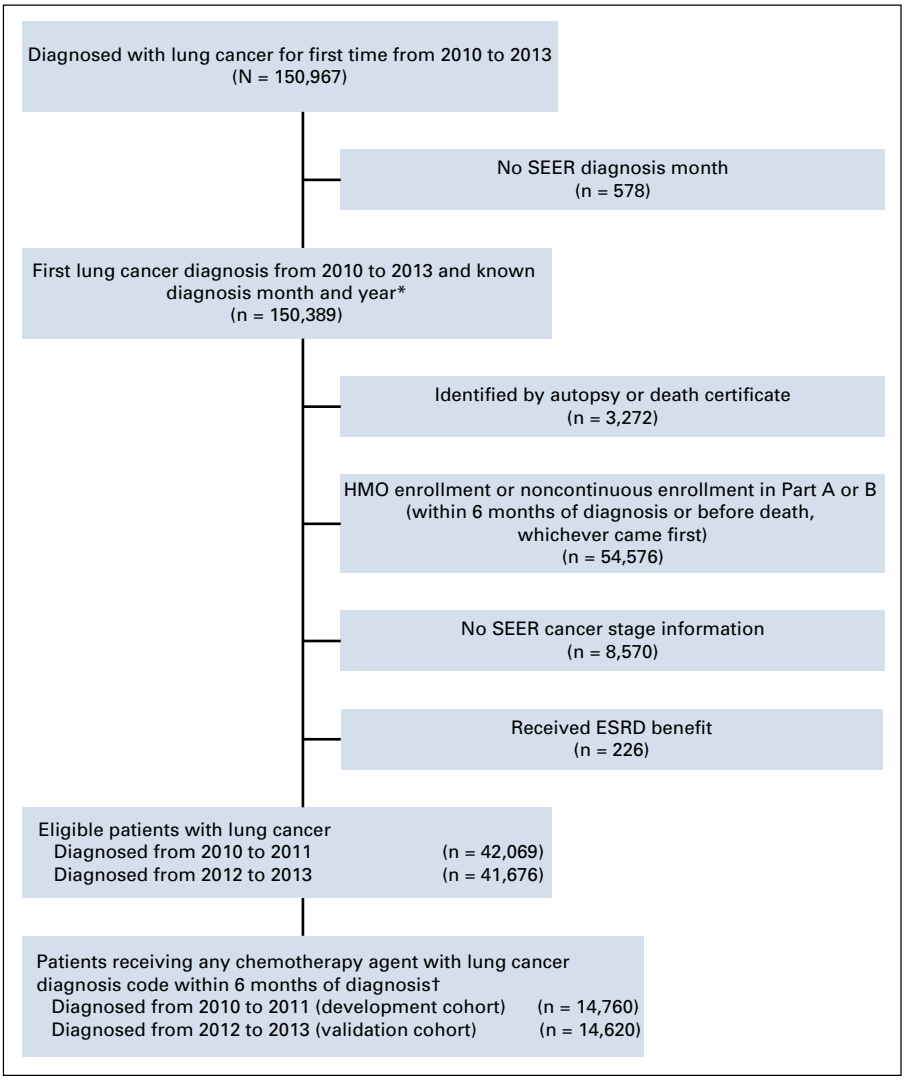

FIG A1. Selection of patient cohorts. ESRD, end-stage renal disease; HMO, health maintenance organization. (*) Diagnosis date set as 15th day of diagnosis month. (†) Excluding 29 patients because of missing zip code of residence.

TABLE A1. Diagnosis, Chemotherapy, and Procedure Codes for Lung Cancer

Code
Lung cancer ICD-9 diagnosis codes
162.0, 162.2, 162.3, 162.4, 162.5, 162.8, 162.9
HCPCS codes for chemotherapy
Cisplatin: J9060, J9062
Carboplatin: J9045
Paclitaxel: J9265, J9267
Docetaxel: J9170, J9171
Paclitaxel, albumin bound: J9264
Pemetrexed: J9305
Gemcitabine: J9201
Vinorelbine: J9390
Bevacizumab: J9035
Etoposide: J8560, J9181, J9182, WW030, WW031, WW032
Irinotecan: J9206
Topotecan: J8705, J9350, J9351
Gefitinib: J8565
Trastuzumab: J9355
CPT and ICD-9 procedure codes for lung resection surgery
Pneumonectomy: 32440, 32442, 32445, 32.50, 32.59
Lobectomy: 32480, 32482, 32484, 32486, 32663, 32.41, 32.49, 32.6
Wedge or segmental resection: 32500, 32505, 32506, 32507, 32657, 32666, 32667, 32668, 32.30, 32.39
CPT codes for radiotherapy
77402, 77403, 77404, 77406, 77407, 77408, 77409, 77411, 77412, 77413, 77414, 77416, 77418, 77522, 77523, 77525, 0073T, 77371, 77372, 77432

NOTE. Healthcare Common Procedure Coding System (HCPCS) codes for ramucirumab, necitumumab, nivolumab, pembrolizumab, atezolizumab, and durvalumab had not yet been issued in 2014 (the last year of claims data used in this analysis).

Abbreviations: CPT, Common Procedural Terminology; ICD-9, International Classification of Diseases, Ninth Revision.

TABLE A2. Descriptive Statistics of Candidate Variables for Machine-Learning Algorithms: Development Cohort (n = 14,760)

Variable	No. (%) of Binary Variables	Continuous Variables				
		Mean	SD	Median	25th	75th
Demographic						
Age at diagnosis, years		72.1	7.7	72.0	67.0	77.0
Male sex	8,062 (54.6)					
Zip code–level median household income, $		60,619	28,580	54,260	—	—
Zip code–level not graduating high school, %		20.2	12.9	17.2	10.1	28.1
Zip code–level residents living below poverty level, %		11.7	9.5	8.8	4.6	16.1
Race/ethnicity						
Black	1,298 (8.8)					
Hispanic	626 (4.2)					
Asian	583 (3.9)					
Other	45 (0.3)					
Region of residence						
Northeast	2,999 (20.3)					
Midwest	2,004 (13.6)					
West	5,049 (34.2)					
South	4,708 (31.9)					
Diagnosis code[a]						
Count of lung cancer diagnosis codes (162.x)						
Rate of 162.0 (malignant neoplasm of trachea)		0.1	1.9	0.0	0.0	0.0
Rate of 162.2 (malignant neoplasm of main bronchus)		1.4	7.4	0.0	0.0	0.0
Rate of 162.3 (malignant neoplasm of upper lobe, bronchus, or lung)		14.6	25.6	0.0	0.0	14.3
Rate of 162.4 (malignant neoplasm of middle lobe, bronchus, or lung)		2.2	11.0	0.0	0.0	0.0
Rate of 162.5 (malignant neoplasm of lower lobe, bronchus, or lung)		6.8	18.3	0.0	0.0	3.3
Rate of 162.8 (malignant neoplasm of other parts of bronchus or lung)		8.9	21.1	0.0	0.0	4.5
Rate of 162.9 (malignant neoplasm of bronchus and lung, unspecified site)		65.9	33.8	81.5	34.1	94.7
Secondary malignancy code[b]						
Count of all secondary malignant codes (196 series + 197 series + 198 series + 199.0)		4.2	7.4	1.0	0.0	5.0
Any secondary malignancy codes[b]	8,498 (57.6)					
Rate of 196 series (secondary and unspecified malignant neoplasm of lymph nodes)		12.8	30.2	0.0	0.0	0.0
Rate of 197 series (secondary malignant neoplasm of respiratory and digestive systems)		18.1	33.5	0.0	0.0	20.0
Rate of 198 series (secondary malignant neoplasm of other specified sites)		26.1	39.6	0.0	0.0	54.5
Rate of 199.0 (disseminated malignant neoplasm without specification of site)		0.6	6.1	0.0	0.0	0.0
Outpatient, inpatient, and critical care[c]						
Count of outpatient E&M claims		12.0	6.1	11.0	8.0	15.0
Any outpatient E&M claims	14,686 (99.5)					
Count of inpatient E&M claims		4.6	6.7	2.0	0.0	6.0
Any inpatient E&M claims	9,840 (66.7)					

(Continued on following page)

TABLE A2. Descriptive Statistics of Candidate Variables for Machine-Learning Algorithms: Development Cohort (n = 14,760) (Continued)

Variable	No. (%) of Binary Variables	Continuous Variables				
		Mean	SD	Median	25th	75th
Count of critical care/advanced life support claims		0.2	0.8	0.0	0.0	0.0
Any critical care/advanced life support claims	1,264 (8.6)					
Count of hospital discharges		1.2	1.2	1.0	0.0	2.0
Any hospital discharges	9,934 (67.3)					
Count of Part B chemotherapy dates (dates with ≥ 1 chemotherapy claims)		6.0	3.9	5.0	3.0	8.0
Any Part B chemotherapy dates	14,138 (95.8)					
Chemotherapy[d]						
Cisplatin, No. of claims		0.7	2.1	0.0	0.0	0.0
Any	2,548 (17.3)					
Carboplatin, No. of claims		3.1	2.7	3.0	0.0	5.0
Any	10,839 (73.4)					
Paclitaxel, No. of claims		1.8	2.9	0.0	0.0	3.0
Any	5,198 (35.2)					
Docetaxel, No. of claims		0.3	1.3	0.0	0.0	0.0
Any	1,275 (8.6)					
Pemetrexed, No. of claims		0.7	1.5	0.0	0.0	0.0
Any	3,047 (20.6)					
Gemcitabine, No. of claims		0.5	1.6	0.0	0.0	0.0
Any	1,371 (9.3)					
Vinorelbine, No. of claims		0.2	1.2	0.0	0.0	0.0
Any	515 (3.5)					
Bevacizumab, No. of claims		0.3	1.1	0.0	0.0	0.0
Any	1,370 (9.3)					
Etoposide, No. of claims		2.3	4.5	0.0	0.0	0.0
Any	3,663 (24.8)					
Irinotecan, No. of claims		0.0	0.5	0.0	0.0	0.0
Any	146 (1.0)					
Topotecan, No. of claims		0.0	0.4	0.0	0.0	0.0
Any	66 (0.4)					
Trastuzumab, No. of claims		0.0	0.1	0.0	0.0	0.0
Any	< 11 (< 1)[e]					
J code unclassified chemotherapy drug, No. of claims		0.0	0.2	0.0	0.0	0.0
Any	28 (0.2)					
Surgery and procedure[f]						
Lung resection surgery within ± 3 months of first chemotherapy						
None	12,749 (86.4)					
Lobectomy lung resection type	1,319 (8.9)					
Pneumonectomy lung resection type	103 (0.7)					
Segmental lung resection type	589 (4.0)					
Radiotherapy[g]						
Count of radiation fractions delivered (radiation treatment delivery codes, excluding stereotactic)		10.8	13.9	0.0	0.0	23.0

(Continued on following page)

TABLE A2. Descriptive Statistics of Candidate Variables for Machine-Learning Algorithms: Development Cohort (n = 14,760) (Continued)

Variable	No. (%) of Binary Variables	Continuous Variables				
		Mean	SD	Median	25th	75th
Any radiation fractions delivered	6,913 (46.8)					
Count of nonbrain stereotactic radiation delivery codes		0.0	0.3	0.0	0.0	0.0
Any nonbrain stereotactic radiation delivered	176 (1.2)					
Count of brain stereotactic radiation delivery codes		0.0	0.2	0.0	0.0	0.0
Any brain stereotactic radiation delivered	277 (1.9)					
Count of nonstereotactic radiation fractions delivered within 60 days of first radiation treatment		10.9	13.7	0.0	0.0	25.0
Any nonstereotactic radiation delivered within 60 days of first radiation treatment	6,913 (46.8)					
Diagnosis code associated with first nonstereotactic radiation delivery code (among patients receiving radiotherapy)						
Lung cancer	6,007 (40.7)					
Secondary malignancy	687 (4.7)					
Other	219 (1.5)					
No radiotherapy	7,847 (53.1)					
Was first radiation delivery code (any) before surgery (among patients receiving radiotherapy)?						
No radiotherapy	7,847 (53.1)					
Yes	96 (0.7)					
No, first radiation treatment after surgery	303 (2.1)					
No, no surgery	6,514 (44.1)					
Comorbidity[h]						
Alzheimer's or other dementia	995 (6.7)					
Acute MI	791 (5.4)					
Ischemic heart disease	8,481 (57.5)					
Stroke/TIA	2,058 (13.9)					
Atrial fibrillation	2,336 (15.8)					
Heart failure	4,557 (30.9)					
Hypertension	12,308 (83.4)					
Hyperlipidemia	11,431 (77.4)					
Diabetes	5,541 (37.5)					
Asthma	2,637 (17.9)					
Chronic obstructive pulmonary disease	9,974 (67.6)					
Depression	4,032 (27.3)					
Chronic kidney disease	3,461 (23.4)					
Hip/pelvic fracture	291 (2.0)					
Decision points of algorithm						
Receipt of lung cancer chemotherapy	14,678 (99.4)					
Use of bevacizumab or stereotactic cranial irradiation within 3 months of initial lung cancer chemotherapy	1,647 (11.2)					
Lung resection surgery in 3 months before initial lung cancer chemotherapy	1,702 (11.5)					
Radiation, ≥ 20 fractions beginning < 7 days before initial lung chemotherapy	3,365 (22.8)					

(Continued on following page)

TABLE A2. Descriptive Statistics of Candidate Variables for Machine-Learning Algorithms: Development Cohort (n = 14,760) (Continued)

Variable	No. (%) of Binary Variables	Continuous Variables				
		Mean	SD	Median	25th	75th
Small-cell chemotherapy agents and platinums only (in 3 months after first lung cancer chemotherapy)	3,530 (23.9)					
Targeted agents (in 3 months after first lung cancer chemotherapy)	718 (4.9)					

Abbreviations: E&M, evaluation and management; MI, myocardial infarction; SD, standard deviation; TIA, transient ischemic attack.

[a]Percentage of all International Classification of Diseases, Ninth Revision (ICD-9) lung cancer diagnosis codes (162.x) that are the specified diagnosis code, from E&M claims (inpatient, outpatient, critical care), MEDPAR hospitalizations, and chemotherapy claims ± 3 months of first chemotherapy (claim level, using claims with lung cancer diagnosis).

[b]Percentage of diagnosis codes 162.x, 196.x, 197.x, 198.x, and 199.0 that are in the specified diagnosis code series.

[c]Outpatient, inpatient, and critical care visits within ± 3 months of first chemotherapy (not restricted to lung cancer diagnosis code).

[d]No. of chemotherapy claims within ± 3 months of first chemotherapy date.

[e]Sample sizes < 11 are suppressed.

[f]Procedures within ± 3 months of first chemotherapy date.

[g]Radiation delivered within ± 3 months of first chemotherapy date.

[h]First occurrence before first chemotherapy date.

TABLE A3. Descriptive Statistics of Candidate Variables for Machine-Learning Algorithms: Validation Cohort (n = 14,620)

Variable	No. (%) of Binary Variables	Continuous Variables				
		Mean	SD	Median	25th	75th
Demographic						
Age at diagnosis, years		71.9	7.7	72.0	67.0	77.0
Male sex	7,779 (53.2)					
Zip code–level median household income, $		60,766	28,933	54,570	40,166	74,946
Zip code–level not graduating high school, %		14.8	10.8	12.1	6.7	20.7
Zip code–level residents living below poverty level, %		14.3	10.8	11.5	6.2	19.8
Race/ethnicity						
Black	1,307 (8.9)					
Hispanic	642 (4.4)					
Asian	706 (4.8)					
Other	66 (0.5)					
Region of residence						
Northeast	2,952 (20.2)					
Midwest	1,923 (13.2)					
West	5,042 (34.5)					
South	4,703 (32.2)					
Diagnosis code*						
Count of lung cancer diagnosis codes (162.x)						
Rate of 162.0 (malignant neoplasm of trachea)		0.1	1.7	0.0	0.0	0.0
Rate of 162.2 (malignant neoplasm of main bronchus)		1.5	8.0	0.0	0.0	0.0
Rate of 162.3 (malignant neoplasm of upper lobe, bronchus, or lung)		15.7	26.4	2.4	0.0	16.7
Rate of 162.4 (malignant neoplasm of middle lobe, bronchus, or lung)		2.0	10.2	0.0	0.0	0.0
Rate of 162.5 (malignant neoplasm of lower lobe, bronchus, or lung)		7.7	19.6	0.0	0.0	3.8
Rate of 162.8 (malignant neoplasm of other parts of bronchus or lung)		7.4	18.0	0.0	0.0	4.0
Rate of 162.9 (malignant neoplasm of bronchus and lung, unspecified site)		65.6	32.9	78.9	35.7	94.1
Secondary malignancy code†						
Count of all secondary malignant codes (196 series + 197 series + 198 series + 199.0)		4.6	7.7	1.0	0.0	6.0
Any secondary malignancy codes	8,686 (59.4)					
Rate of 196 series (secondary and unspecified malignant neoplasm of lymph nodes)		14.1	31.4	0.0	0.0	0.0
Rate of 197 series (secondary malignant neoplasm of respiratory and digestive systems)		17.7	32.7	0.0	0.0	20.0
Rate of 198 series (secondary malignant neoplasm of other specified sites)		27.2	39.7	0.0	0.0	60.0
Rate of 199.0 (disseminated malignant neoplasm without specification of site)		0.5	5.3	0.0	0.0	0.0
Outpatient, inpatient, and critical care‡						
Count of outpatient E&M claims		12.1	6.2	11.0	8.0	15.0
Any outpatient E&M claims	14,568 (99.6)					
Count of inpatient E&M claims		4.6	6.9	2.0	0.0	6.0
Any inpatient E&M claims	9,574 (65.5)					

(Continued on following page)

TABLE A3. Descriptive Statistics of Candidate Variables for Machine-Learning Algorithms: Validation Cohort (n = 14,620) (Continued)

Variable	No. (%) of Binary Variables	Continuous Variables				
		Mean	SD	Median	25th	75th
Count of critical care/advanced life support claims		0.2	0.7	0.0	0.0	0.0
Any critical care/advanced life support claims	1,308 (8.9)					
Count of hospital discharges		1.1	1.2	1.0	0.0	2.0
Any hospital discharges	9,534 (65.2)					
Count of Part B chemotherapy dates (dates with ≥ 1 chemotherapy claims)		5.9	3.9	5.0	3.0	8.0
Any Part B chemotherapy dates	13,820 (94.5)					
Chemotherapy§						
Cisplatin, No. of claims		0.7	1.8	0.0	0.0	0.0
Any	2,413 (16.5)					
Carboplatin, No. of claims		3.1	2.6	3.0	0.0	5.0
Any	10,841 (74.2)					
Paclitaxel, No. of claims		1.8	3.0	0.0	0.0	3.0
Any	5,062 (34.6)					
Docetaxel, No. of claims		0.2	1.0	0.0	0.0	0.0
Any	798 (5.5)					
Pemetrexed, No. of claims		0.8	1.6	0.0	0.0	0.0
Any	3,555 (24.3)					
Gemcitabine, No. of claims		0.3	1.4	0.0	0.0	0.0
Any	999 (6.8)					
Vinorelbine, No. of claims		0.1	1.0	0.0	0.0	0.0
Any	317 (2.2)					
Bevacizumab, No. of claims		0.3	1.0	0.0	0.0	0.0
Any	1,289 (8.8)					
Etoposide, No. of claims		2.4	4.6	0.0	0.0	2.0
Any	3,719 (25.4)					
Irinotecan, No. of claims		0.0	0.4	0.0	0.0	0.0
Any	97 (0.7)					
Topotecan, No. of claims		0.0	0.4	0.0	0.0	0.0
Any	39 (0.3)					
Trastuzumab, No. of claims		0.0	0.1	0.0	0.0	0.0
Any	16 (0.1)					
J code unclassified chemotherapy drug, No. of claims		0.0	0.1	0.0	0.0	0.0
Any	18 (0.1)					
Surgery and procedure‖						
Lung resection surgery within ± 3 months of first chemotherapy						
None	12,745 (87.2)					
Lobectomy lung resection type	1,250 (8.5)					
Pneumonectomy lung resection type	104 (0.7)					
Segmental lung resection type	521 (3.6)					
Radiotherapy¶						
Count of radiation fractions delivered (radiation treatment delivery codes, excluding stereotactic)		10.5	13.6	0.0	0.0	22.0

(Continued on following page)

TABLE A3. Descriptive Statistics of Candidate Variables for Machine-Learning Algorithms: Validation Cohort (n = 14,620) (Continued)

Variable	No. (%) of Binary Variables	Continuous Variables				
		Mean	SD	Median	25th	75th
Any radiation fractions delivered	6,851 (46.9)					
Count of nonbrain stereotactic radiation delivery codes		0.0	0.3	0.0	0.0	0.0
Any nonbrain stereotactic radiation delivered	250 (1.7)					
Count of brain stereotactic radiation delivery codes		0.0	0.2	0.0	0.0	0.0
Any brain stereotactic radiation delivered	330 (2.3)					
Count of nonstereotactic radiation fractions delivered within 60 days of first radiation treatment		10.6	13.5	0.0	0.0	24.0
Any nonstereotactic radiation delivered within 60 days of first radiation treatment	6,851 (46.9)					
Diagnosis code associated with first nonstereotactic radiation delivery code (among patients receiving radiotherapy)						
Lung cancer	5,927 (40.5)					
Secondary malignancy	707 (4.8)					
Other	217 (1.5)					
No radiotherapy	7,769 (53.1)					
Was first radiation delivery code (any) before surgery (among patients receiving radiotherapy)?						
No radiotherapy	7,769 (53.1)					
Yes	90 (0.6)					
No, first radiation treatment after surgery	258 (1.8)					
No, no surgery	6,503 (44.5)					
Comorbidity#						
Alzheimer's or other dementia	1,023 (7.0)					
Acute MI	728 (5.0)					
Ischemic heart disease	8,245 (56.4)					
Stroke/TIA	2,024 (13.8)					
Atrial fibrillation	2,251 (15.4)					
Heart failure	4,410 (30.2)					
Hypertension	12,258 (83.8)					
Hyperlipidemia	11,516 (78.8)					
Diabetes	5,693 (38.9)					
Asthma	2,850 (19.5)					
Chronic obstructive pulmonary disease	9,772 (66.8)					
Depression	4,505 (30.8)					
Chronic kidney disease	3,740 (25.6)					
Hip/pelvic fracture	335 (2.3)					
Decision points of algorithm						
Receipt of lung cancer chemotherapy	14,519 (99.3)					
Use of bevacizumab or stereotactic cranial irradiation within 3 months of initial lung cancer chemotherapy	1,610 (11.0)					
Lung resection surgery in 3 months before initial lung cancer chemotherapy	1,587 (10.9)					
Radiation, ≥ 20 fractions beginning < 7 days before initial lung chemotherapy	3,344 (22.9)					

(Continued on following page)

TABLE A3. Descriptive Statistics of Candidate Variables for Machine-Learning Algorithms: Validation Cohort (n = 14,620) (Continued)

Variable	No. (%) of Binary Variables	Continuous Variables				
		Mean	SD	Median	25th	75th
Small-cell chemotherapy agents and platinums only (in 3 months after first lung cancer chemotherapy)	3,621 (24.8)					
Targeted agents (in 3 months after first lung cancer chemotherapy)	907 (6.2)					

Abbreviations: E&M, evaluation and management; MI, myocardial infarction; SD, standard deviation; TIA, transient ischemic attack.

*Percentage of all International Classification of Diseases, Ninth Revision (ICD-9) lung cancer diagnosis codes (162.x) that are the specified diagnosis code, from E&M claims (inpatient, outpatient, critical care), MEDPAR hospitalizations, and chemotherapy claims ± 3 months of first chemotherapy (claim level, using claims with lung cancer diagnosis).

†Percentage of diagnosis codes 162.x, 196.x, 197.x, 198.x, and 199.0 that are in the specified diagnosis code series.

‡Outpatient, inpatient, and critical care visits within ± 3 months of first chemotherapy (not restricted to lung cancer diagnosis code).

§No. of chemotherapy claims within ± 3 months of first chemotherapy date.

‖Procedures within ± 3 months of first chemotherapy date.

¶Radiation delivered within ± 3 months of first chemotherapy date.

#First occurrence before first chemotherapy date.