Clinical Epidemiology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Development and Validation of Coding Algorithms to Identify Patients with Incident Non-Small Cell Lung Cancer in United States Healthcare Claims Data

Julie Beyrer[1,*], David R Nelson[1,*], Kristin M Sheffield[1,*], Yu-Jing Huang[1,*], Yiu-Keung Lau[1,*], Ana L Hincapie[2,*]

[1]Eli Lilly and Company, Indianapolis, IN, USA; [2]University of Cincinnati James L. Winkle College of Pharmacy, Cincinnati, OH, USA

*These authors contributed equally to this work

Correspondence: Julie Beyrer, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN, 46285, USA, Tel +1 317 651 8236, Email beyrerj@lilly.com

Purpose: We sought to develop and validate an incident non-small cell lung cancer (NSCLC) algorithm for United States (US) healthcare claims data. Diagnoses and procedures, but not medications, were incorporated to support longer-term relevance and reliability.

Methods: Patients with newly diagnosed NSCLC per Surveillance, Epidemiology, and End Results (SEER) served as cases. Controls included newly diagnosed small-cell lung cancer and other lung cancers, and two 5% random samples for other cancer and without cancer. Algorithms derived from logistic regression and machine learning methods used the entire sample (Approach A) or started with a previous algorithm for those with lung cancer (Approach B). Sensitivity, specificity, positive predictive values (PPV), negative predictive values, and F-scores (compared for 1000 bootstrap samples) were calculated. Misclassification was evaluated by calculating the odds of selection by the algorithm among true positives and true negatives.

Results: The best performing algorithm utilized neural networks (Approach B). A 10-variable point-score algorithm was derived from logistic regression (Approach B); sensitivity was 77.69% and PPV = 67.61% (F-score = 72.30%). This algorithm was less sensitive for patients ≥80 years old, with Medicare follow-up time <3 months, or missing SEER data on stage, laterality, or site and less specific for patients with SEER primary site of main bronchus, SEER summary stage 2000 regional by direct extension only, or pre-index chronic pulmonary disease.

Conclusion: Our study developed and validated a practical, 10-variable, point-based algorithm for identifying incident NSCLC cases in a US claims database based on a previously validated incident lung cancer algorithm.

Keywords: algorithm, machine learning, medicare claims, non-small cell lung cancer, positive predictive value, sensitivity, validation

Introduction

Validated algorithms are essential research tools for identifying patient cohorts, exposures, key covariates, and outcomes in real-world data.[1] The accurate identification of patient diagnoses in real-world data sources, such as administrative healthcare databases, is essential for learning about patient disease experiences. For example, validated disease cohort and outcome algorithms can support earlier diagnosis of disease or better screening efforts (eg, by enabling research on predictors of those diagnoses). This is an important consideration for diseases that are currently diagnosed after the disease has progressed to a more serious or advanced stage and the prognosis is poor. Lung cancers are often diagnosed at an advanced stage,[2] despite recommendations by the United States (US) Preventive Services Task Force for low-dose computed tomography (LDCT) in at-risk adults.[3] Lung cancer is a heterogenous disease. Most lung cancers in the US are non-small cell lung cancer (NSCLC; 80% to 85%), which originate in different types of cells and have a different prognosis and prescribed treatments than other types of lung cancer.[4] NSCLC includes the main subtypes of adenocarcinoma, squamous cell carcinoma, and large cell carcinoma, which are classed together as NSCLC because of the

Received: 14 September 2022
Accepted: 23 December 2022
Published: 12 January 2023

similarity of their treatment and prognoses. In contrast, small cell lung cancer (SCLC) tends to grow and spread faster than NSCLC. Consequently, the ability to differentiate NSCLC from other types of lung cancers in real-world data sources is important to enable research on NSCLC diagnosis, comorbidities, treatment patterns, adverse effects, prognosis, healthcare costs and resource utilization.

However, directly identifying NSCLC in administrative healthcare databases in the US is not feasible. The diagnostic coding system used in these databases (the International Classification of Diseases [ICD], Clinical Modification) categorizes cancers according to site of origin rather than pathologic characteristics and it cannot be used to differentiate NSCLC from other types of lung cancer. Other healthcare practices contribute to disease misclassification in these administrative healthcare databases, such as differing interpretations of coding guidelines by medical coders,[5,6] partial and misclassified clinical data, and incomplete claims that are not paid on a fee-for-service basis.[7]

A linked data source (eg, registry data source linked with claims, such as the Surveillance, Epidemiology, and End Results [SEER]-Medicare database) can supplement clinical details in medical claims. When linked data sources are not available, validated claims-based algorithms provide an alternative for identifying clinical conditions in claims. Various analytical methods for building algorithms, such as single classification trees and random forests, can be used to identify patients with cancer.[8–10]

An algorithm to identify patients with NSCLC in administrative claims databases in the US was developed and validated by Turner and colleagues.[11] However, the algorithm criteria included medications for treating NSCLC, which may not be stable indicators over time as the NSCLC treatment landscape is quickly evolving. In some cases, algorithms that include medication(s) may not be generalizable and re-useable in other healthcare administrative data sources if data on the medication(s) are systemically missing. Medication data may be systemically missing if the medication(s) are not covered by the health plan's formulary or if a medication-specific code does not exist during the study timeframe. For example, Healthcare Common Procedure Codes (HCPC) are submitted by healthcare providers on medical claims to obtain reimbursement for medications administered; data may be missing for HCPC-coded medications in some medical claims datasets if those data are pulled before the medical claim has been processed (HCPC-coded medications in medical claims take longer for health plans to process or adjudicate than medications submitted for reimbursement on pharmacy claims and thus may not appear in the data source). The delayed issuance and effective dates of HCPC codes in the US have also historically limited a researcher's ability to identify medications in medical claims data sources during certain periods (approximately one to two years after a medication's approval by the United States Food and Drug Administration), although recent improvements to the HCPCS coding application process may offer shorter timeframes between new medication approval and issuance of a HCPC.[12] Algorithms that use medications to identify cases (patients with disease) are also not ideal because they may preclude the study of real-world treatment patterns (if the cohort has been defined by treatment) or may result in selection bias. For example, older patients with NSCLC or those with comorbidities may be less likely to receive treatment[13,14] and therefore systematically being excluded from the study.

The objective of our study was to develop and validate an algorithm for identifying incident NSCLC cases in US healthcare claims data. The development process avoided the use of medications and incorporated diagnostic and procedural codes and other concepts that are less likely to change over time or be missing in administrative healthcare data. This supports the long-term relevance and reliability of the algorithm.

Materials and Methods

Data Sources

We used the SEER registry linked with Medicare claims data (2004 to 2012). Full details of the data source used in this study were published previously.[15] The protocol was reviewed and considered exempt by the Quorum Review institutional review board prior to approval by the National Cancer Institute for SEER-Medicare data use.

Study Populations

Patients with newly diagnosed (incident) NSCLC per SEER served as cases. Controls included newly diagnosed (incident) small-cell lung cancer (SCLC) and other lung cancers, a 5% random sample of patients with other cancers (the majority were colorectal, female breast, and prostate cancers), and a 5% random sample of individuals without

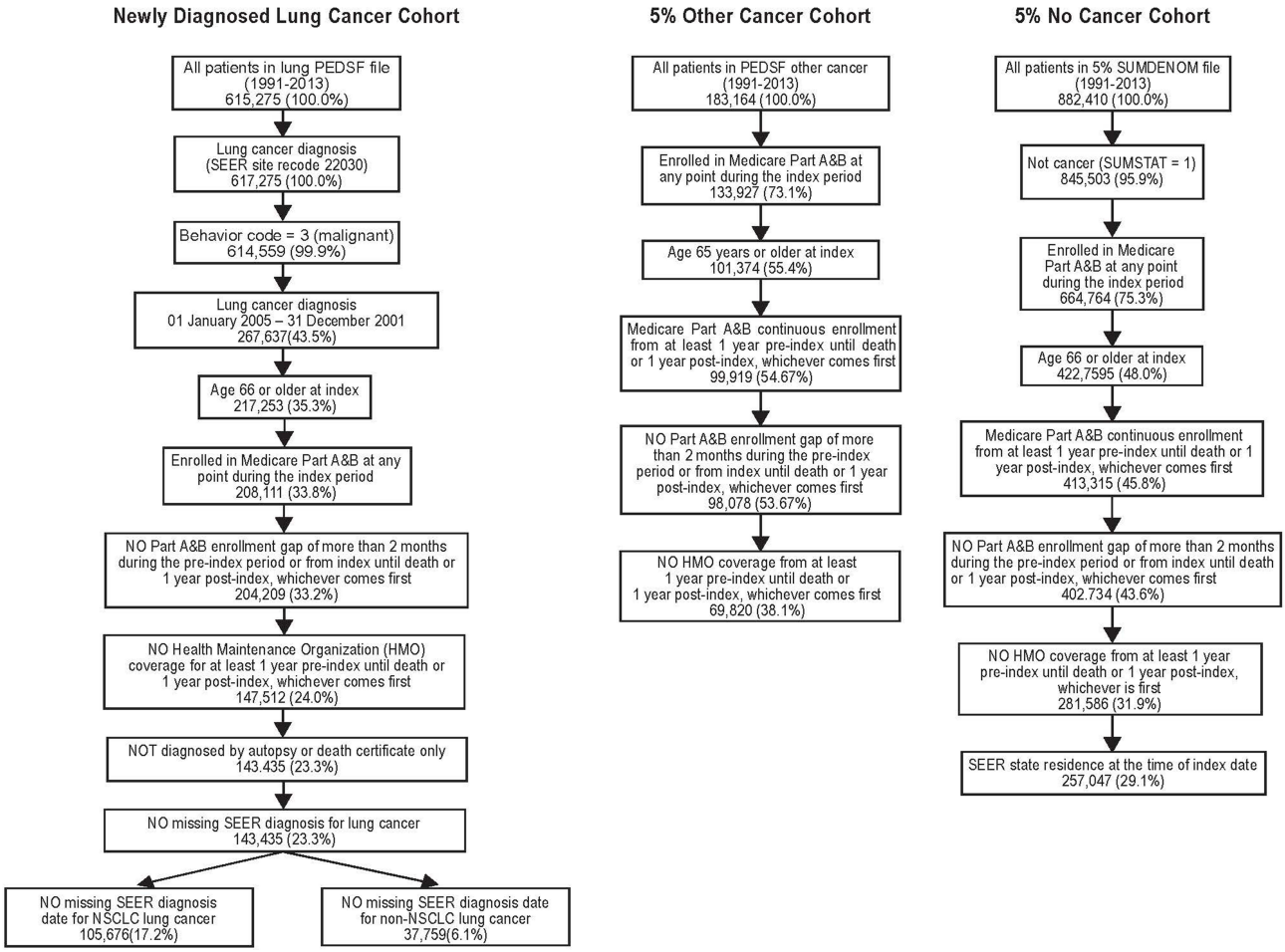

Figure 1 Attrition of cohorts of lung cancer, other cancer, and non-cancer cohorts from SEER-Medicare.
Notes: A primary diagnosis of NSCLC included SEER histology codes in these ranges: 8003–8004, 8012–8015, 8021–8022, 8030–8035, 8046, 8050–8052, 8070–8076, 8078, 8082–8084, 8090, 8094, 8120, 8123, 8140–8141, 8143–8145, 8147, 8190, 8200–8201, 8211, 8240–8241, 8243–8246, 8249–8255, 8260, 8290, 8310, 8320, 8323, 8333, 8401, 8430, 8440, 8470–8471, 8480–8481, 8490, 8503, 8507, 8525, 8550, 8560, 8562, 8570–8572, 8574–8576.
Abbreviations: HMO, health maintenance organization; NSCLC, non-small cell lung cancer; PEDSF, Patient Entitlement and Diagnosis Summary; SEER, Surveillance, Epidemiology, and End Results.

cancer. Patients in the incident NSCLC cohort were required to have a primary diagnosis of NSCLC (SEER histology codes are shown in Figure 1) and behavior code = 3 (malignant) during the index period (January 1, 2005 through December 31, 2011; the full study period was January 1, 2004 through December 31, 2012). Medicare coverage typically starts at age 65, so patients were required to be 66 years or older at SEER initial lung cancer diagnosis date to help ensure ≥1 year of pre-index diagnosis data would be available for constructing the algorithm.[16]

Patients who were diagnosed by autopsy or death certificate only or who were missing a SEER diagnosis date for lung cancer were excluded. Evidence of Medicare Parts A&B enrollment was also required from ≥1 year pre-index until death or one-year post index, whichever was first. The initial SEER documented date (month/year) of diagnosis was utilized for inclusion of patients in the incident NSCLC cohort. The same criteria that applied to the initial SEER lung cancer diagnosis date were used to identify incident SCLC and other lung cancer case controls. A randomly assigned date for each individual between January 1, 2005 and death or end of the index period (December 31, 2011) was utilized to select controls for the 5% other cancers and 5% non-cancer cohorts.

Model Building and Validation Subsets

The full dataset was randomly split into model building and validation subsets of 50% each (if total was an odd number, one more observation was included in the model building subset), with selection stratified by cohort (lung cancer, other

cancer, non-cancer). The model building subset was used to extensively build models (eg, to identify significant interactions using multivariate adaptive regression splines methods) and construct models/algorithms. The model validation subset was reserved to obtain unbiased estimates of algorithm performance. Candidate variables were chosen based on clinical practice guidelines (eg, National Comprehensive Cancer Network)[17] and observed frequencies of variables for cases and controls, as well as consultation with experts. The algorithms were constructed using the initial ICD, Ninth Revision, Clinical Modification (ICD-9-CM) diagnosis code for lung cancer (162.2–162.9) to create one-year pre-index and post-index periods for NSCLC and other lung cancer cohorts. The randomly assigned Medicare date was used for the other cancer and non-cancer individuals. We searched for candidate variables in the pre- and post-index periods in Medicare claims. Based on the algorithms that performed best in our previous study,[15] the following models were explored in this study: a single logistic regression model, a single logistic regression model with interactions, gradient boosting, and neural networks. The multilayer perceptron neural network had two hidden layers in the network and two neurons in each hidden layer because five hidden nodes did not improve the F-score in the model building set. The maximum number of iterations (weight adjustments) for the optimizer to make before terminating was set at 1000, and every fourth observation was a model tuning observation.

Our reference standard was the type of lung cancer as identified by SEER. However, the National Cancer Institute estimates that approximately 4.5% of SEER data could not be linked to Medicare claims (personal communication); therefore, some patients with lung cancer could have inadvertently been delivered in the 5% random non-cancer sample. We observed that 688 (0.3%) of the reference standard "non-cancer" patients had 34+ days of ICD-9-CM codes for lung cancer in Medicare claims. We re-categorized some of these individuals as lung cancer cases, similar to that done by Nattinger and colleagues,[16] which resulted in 193 patients being re-categorized from non-cancer to lung cancer in this study. Although these patients were included as lung cancer cases in the previous study,[15] they were excluded from this study. This was necessary since they were not in the SEER registry and did not have information on lung cancer type. These patients could also not be included as controls since they were likely to be cases.

Algorithm Development

Two approaches were used to develop the NSCLC algorithms. The algorithm in Approach A was built based on the entire sample, designating patients with NSCLC as cases and patients with other types of lung cancer, other cancers, and no cancer as controls. Approach B was a two-step process that applied a previously validated lung cancer algorithm[15] to identify patients with lung cancer in the sample and then designated patients as cases and controls as outlined in Approach A.

We also explored the impact of using or not using medication indicators (eg, exclusion of chemotherapies typically indicated for SCLC during this time period [cyclophosphamide, doxorubicin, irinotecan, vincristine, or topotecan]) for identifying NSCLC.

Statistical Analysis

Sensitivity, specificity, positive predictive values (PPV), negative predictive values (NPV), and F-scores (the product of sensitivity and PPV multiplied by two and divided by the sum of sensitivity and PPV) were calculated in the model building and validation subsets. Due to over-representation (enriched prevalence) of patients with lung cancer in the study sample, PPVs and NPVs were calculated using Bayes' theorem.[15,16] Here, Bayes' theorem values were: for NSCLC (0.29%; or 73.7%, based on SEER, of the 0.40% lung cancer rate), other lung cancer (the remaining 0.11%), other cancer (7.3%), and non-cancer (92.2%) for Approach A and Approach B.

Four algorithm types were evaluated in the validation subset, and logistic regression models were repeated to explore models including and excluding medications/chemotherapy variables. A simplified point score was created by reducing the number of variables in the logistic model feature selection,[18] then with forward selection to stop at 10 variables. The coefficients of the 10 variable model were converted to points by dividing each by the absolute value of the smallest coefficient and rounding to the nearest integer, as described in risk scores, such as Framingham[19] and others.[20] Bootstrap samples of the validation dataset were selected 1000 times for one-tailed p-values to test each algorithm vs the point score based on 10 variables.

Univariate and multivariable logistic regression analyses were performed separately within two groups (NSCLC and non-NSCLC patients) to determine the association of variables with sensitivity and specificity. For the univariate analyses, false discovery rate p-values were used to assess significance. A multivariable logistic model was generated by using feature selection of the variables in the univariate analysis, followed by forward model selection. For both the sensitivity and specificity models, forward selection was stopped when the model's area under the receiver operating characteristic curve was within 1% of a larger stepwise model. A version of number needed to treat (NNT) was calculated to indicate how many patients within a subgroup that had reduced either sensitivity or specificity would result in one additional false negative or false positive, respectively. The purpose of the NNTs was to illustrate the impact of these significant variables.

Summary statistics are presented as mean (standard deviation) or percentage. All computations used SAS software version 9.4 (SAS Institute Inc., Cary, NC, USA).

Results

The cohort attrition in the SEER-Medicare dataset is displayed (Figure 1).

The non-NSCLC lung cancer cohort was comprised of individuals with SCLC (n = 16,871 [44.7%]) and other lung cancer (n = 20,888 [55.3%]). Characteristics of the SEER-Medicare cohorts, including a summary of demographics and components of the final point-based score algorithm, are shown in Table 1. Mean cohort ages ranged from 76 to 78 years, and the majority (>80%) of patients were White. Male sex was less frequent in the non-cancer cohort (38%), equally distributed among the other control groups, and slightly more frequent in the NSCLC cohort (53%). A descriptive summary of the cohorts based on all candidate variables used to build the algorithms in this study is presented in S1.

Model Building Subset

The performance results derived from the model building and validation subsets are summarized in Table 2. In the model building subset, using either Approach A or Approach B, the best performing algorithm based on F-scores was identified using neural networks. When using Approach B compared to Approach A, the neural networks model PPV increased from 54.37% to 72.36% and the F-score increased from 65.82% to 74.53%. For Approach B, the remaining models had F-scores ranging from 71.57% to 73.89%, indicating Approach B was superior to Approach A in this algorithm measure. The cut point of ≥67% probability of being NSCLC for the logistic regression model optimized the F-score (F = 81.37%) within the subset of lung cancer algorithm-positive patients utilized in Approach B; this model was used to derive the 10-variable point-based score algorithm, which had a maximum F-score with a cutoff of ≥5 points (the point-score algorithm can be found in Table 3; cut points are shown in S2 and the logistic regression model is detailed in S3).

Model Validation Subset

When these algorithms were applied to new patients in the model validation subset, a similar pattern was observed; the best performing algorithm was the neural network model (Table 2). The PPV in the neural networks model increased from 55.42% to 72.32% and the F-score increased from 66.60% to 74.57% when starting with Approach B compared to Approach A. For Approach B, the remaining models had F-scores ranging from 71.99% to 73.75%. In both the model building and validation subsets, the exploratory models based on medication indicators (eg, exclusion of SCLC medications to identify patients with NSCLC) had slightly increased sensitivity and slightly decreased PPV compared with models that did not include medications. The point-based score algorithm using Approach B performed better than all methods that used Approach A. Within Approach B, five algorithms (excluding those with evidence of cyclophosphamide, doxorubicin, irinotecan, vincristine, or topotecan; logistic regression; and machine learning models) performed significantly better than the point score (all p≤0.01; Table 2).

Algorithm Sensitivity and Specificity by Patient Characteristics

Sensitivity and specificity of the point-based algorithm by selected characteristics from the SEER registry are displayed in Tables 4 and 5. In the multivariable analysis, these characteristics were associated with significantly lower odds of

Table 1 Characteristics of Demographics, Rates of Lung Cancer Codes, and Components of the Algorithm Developed to Identify Cases (Non-Small Cell Lung Cancer; NSCLC) from Controls (Three Other Surveillance, Epidemiology, and End Results-Medicare Cohorts)

Variables	NSCLC (n=105,676)	Non-NSCLC LC (n=37,759)	Non-Cancer (n=257,047)	Other Cancer[b] (n=69,820)
	Mean (SD) or Pct.	Mean (SD) or Pct.	Mean (SD) or Pct.	Mean (SD) or Pct.
Age	76.05 (6.48)	77.95 (7.47)	77.12 (8.01)	77.83 (7.53)
Race/ethnicity				
Unknown	0.09%	0.11%	0.15%	0.11%
White	86.03%	88.06%	82.62%	86.28%
Black	8.28%	7.47%	7.77%	7.39%
Other	1.55%	1.03%	2.49%	1.87%
Asian	2.77%	1.99%	4.14%	2.60%
Hispanic	1.05%	1.04%	2.47%	1.48%
North American Native	0.23%	0.30%	0.37%	0.25%
Male	53.29%[a]	48.67%	38.48%	49.24%
At least 1 LC code	98.23%	94.41%	1.41%	6.24%
Enlargement of lymph nodes – post index	27.44%[a]	22.70%	2.70%	9.21%
10 or more days with LC codes	93.28%[a]	76.47%	0.47%	1.79%
PET scan –post index	55.15%[a]	25.85%	1.30%	10.58%
Pleurisy, pleural effusion, or empyema – post index	55.51%[a]	38.67%	16.06%	23.43%
162.2 malignant neoplasm of main bronchus	14.98%[a]	15.11%	0.06%	0.15%
Presence of ICD-9-CM codes 162.2–162.9 on outpatient claim	77.64%[a]	49.40%	0.45%	2.06%
Lung biopsy pre- or post-	80.49%[a]	48.29%	3.11%	5.89%
Lung resection Pre- or post-	26.65%[a]	2.84%	0.23%	0.61%
CCI – chronic pulmonary disease	73.07%[a]	76.00%	26.35%	27.41%

Notes: Index is the first ICD-9-CM code of 162.2 through 162.9 in the Medicare claims. The pre-index period used in our study was one year. Patients were followed until the earlier of death or one-year post index. Machine readable code lists are available at https://doi.org/10.5281/zenodo.5095308. [a]Selected for point score algorithm to best discriminate NSCLC vs non-NSCLC LC/non-cancer/other cancer. [b]Other Cancer=these were patients in the 5% random sample from SEER areas who were reported to have "other cancer." Most of these were colorectal, female breast, and prostate cancers.

Abbreviations: ICD-9-CM, International Classification of Diseases, Ninth Revision, Clinical Modification; CCI, Charlson Comorbidity Index; LC, lung cancer; NSCLC, non-small cell lung cancer; pct, percent; PET, positron emission tomography; SD, standard deviation; SEER, Surveillance, Epidemiology, and End Results.

cases being selected by the algorithms (ie, lower sensitivity): short follow-up time in Medicare (<3 months); derived American Joint Committee on Cancer (AJCC) stage not applicable; no information concerning laterality; derived AJCC stage IA; age at index ≥80; SEER primary site of lung not otherwise specified; and year of diagnosis (based on SEER) of 2011. The NNT to produce one additional false negative within these subgroups ranged from 3.2 (AJCC stage not applicable) to 11.7 (year 2011).

These characteristics were associated with significantly lower algorithm specificity: chronic pulmonary disease (pre-index Charlson comorbidity), SEER primary site of main bronchus, and SEER summary stage 2000 regional by direct extension only (Table 5). For these subgroups, NNTs ranged from 3.4 (bronchus: main) to 5.9 (chronic pulmonary disease) to produce one additional false positive.

Discussion

The best performing algorithm for identifying incident NSCLC cases was a neural network machine learning model (Approach B). However, machine learning models could not easily be converted to a point-based system for re-use and may pose additional challenges for interpretability,[21,22] so the logistic regression models remained the most practical application. Our logistic regression model was reduced from 77 variables to 10 for the point-score algorithm (Approach B) which was considered the most practical algorithm for general future use. Despite the statistically significant better performance of

Table 2 Comparison of Methods to Discriminate NSCLC from Medicare Claims Built to Maximize Their F-Score on One-Half of the Data (Model Building), and Applied to the Other Half of the Data (Validation)

	Model Building Subset							Testing Models with the Validation Subset							
	NSCLC Sens	Non-Cancer Spec	Other Cancer Spec[a]	Non-NSCLC LC Spec	Bayes' PPV	Bayes' NPV	F-Score	NSCLC Sens	Non-Cancer Spec	Other Cancer Spec	Non-NSCLC LC Spec	Bayes' PPV	Bayes' NPV	F-score	p-value
Approach A – Models Based on Entire Sample:															
A. Logistic regression															
1. Logistic regression after variable selection	79.90	99.87	99.39	66.09	53.62	99.94	64.17	80.09	99.88	99.44	66.13	55.54	99.94	65.59	1.00
2. Logistic regression with interactions (from MARS)	85.45	99.82	99.19	59.24	48.73	99.96	62.06	85.39	99.82	99.30	59.10	49.49	99.96	62.66	1.00
B. Boosted tree	80.34	99.83	99.33	64.91	49.00	99.94	60.87	80.52	99.83	99.35	65.11	49.72	99.94	61.48	1.00
C. Neural networks	83.38	99.87	99.37	63.31	54.37	99.95	65.82	83.43	99.88	99.39	63.39	55.42	99.95	66.60	1.00
Approach B – Models Based on Lung Cancer Positive Score Subgroup Applied to Entire Sample															
A. Logistic regression															
1. Logistic regression after variable selection	73.71	99.97	99.83	68.40	74.07	99.92	73.89	73.97	99.96	99.81	68.28	72.65	99.92	73.31	0.01
2. Logistic regression with interactions (from MARS)	66.43	99.98	99.88	71.76	77.56	99.90	71.57	77.98	99.95	99.77	62.04	69.95	99.93	73.75	<0.001
B. Boosted tree	73.70	99.96	99.81	67.68	72.38	99.92	73.04	74.01	99.96	99.82	67.76	72.12	99.92	73.05	<0.001
C. Neural networks	76.83	99.96	99.77	65.47	72.36	99.93	74.53	76.97	99.96	99.80	65.41	72.32	99.93	74.57	<0.001
EvdCyclo Post Score	77.67	99.96	99.79	60.14	69.98	99.93	73.62	77.97	99.95	99.80	60.14	69.17	99.93	73.31	<0.001
EvdChemo Post Score	76.00	99.96	99.80	56.86	69.42	99.93	72.56	76.28	99.95	99.80	57.19	68.16	99.93	71.99	0.95
NSCLC Score/No Meds	77.34	99.95	99.78	55.22	68.37	99.93	72.58	77.69	99.95	99.79	55.57	67.61	99.93	72.30	–

Notes: [a]Other Cancer=these were patients in the 5% random sample from SEER areas who were reported to have "other cancer." Most of these were colorectal, female breast, and prostate cancers.

Abbreviations: LC, lung cancer; MARS, multivariate adaptive regression splines; med, medication; NPV, negative predictive value; NSCLC, non-small cell lung cancer; PPV, positive predictive value; prob, probability; sens, sensitivity; spec, specificity.

Table 3 Point-Based Algorithm[a] (Based on Logistic Regression Equation) with a Recommended Code Set for International Classification of Diseases, 10th Revision, Clinical Modification

Lung Resection Pre- or Post-Index		**Algorithm Points if ≥ 1 Code Present**
CPT	31640, 32440, 32442, 32445, 32480, 32482, 32484, 32486, 32488, 32491, 32500, 32501, 32503, 32504, 32505, 32506, 32507, 32520, 32522, 32525, 32657, 32663, 32666, 32667, 32668, 32669, 32670, 32671, 0251T, 0252T	+8
ICD-9 procedures	33.20, 33.24, 33.25, 33.26, 33.27, 33.28, 34.20, 34.23, 34.24, 34.25, 34.26, 34.27	
ICD-10[b] procedures	0B534ZZ, 0B538ZZ, 0B544ZZ, 0B548ZZ, 0B554ZZ, 0B558ZZ, 0B564ZZ, 0B568ZZ, 0B574ZZ, 0B578ZZ, 0B584ZZ, 0B588ZZ, 0B594ZZ, 0B598ZZ, 0B5B4ZZ, 0B5B8ZZ, 0BB34ZZ, 0BB38ZZ, 0BB44ZZ, 0BB48ZZ, 0BB54ZZ, 0BB58ZZ, 0BB64ZZ, 0BB68ZZ, 0BB74ZZ, 0BB78ZZ, 0BB84ZZ, 0BB88ZZ, 0BB94ZZ, 0BB98ZZ, 0BBB4ZZ,0BBB8ZZ, 0B530ZZ,0B533ZZ, 0B537ZZ, 0B540ZZ, 0B543ZZ, 0B547ZZ, 0B550ZZ, 0B553ZZ, 0B557ZZ, 0B560ZZ, 0B563ZZ, 0B567ZZ, 0B570ZZ, 0B573ZZ, 0B577ZZ, 0B580ZZ, 0B583ZZ, 0B587ZZ, 0B590ZZ, 0B593ZZ, 0B597ZZ, 0B5B0ZZ, 0B5B3ZZ, 0B5B7ZZ, 0BB30ZZ, 0BB33ZZ, 0BB37ZZ, 0BB40ZZ, 0BB43ZZ, 0BB47ZZ, 0BB50ZZ, 0BB53ZZ, 0BB57ZZ, 0BB60ZZ, 0BB63ZZ, 0BB67ZZ, 0BB70ZZ, 0BB73ZZ, 0BB77ZZ, 0BB80ZZ, 0BB83ZZ, 0BB87ZZ, 0BB90ZZ, 0BB93ZZ, 0BB97ZZ, 0BBB0ZZ, 0BBB3ZZ, 0BBB7ZZ, 0BT30ZZ, 0BT34ZZ, 0BT40ZZ, 0BT44ZZ, 0BT50ZZ, 0BT54ZZ, 0BT60ZZ, 0BT64ZZ, 0BT70ZZ, 0BT74ZZ, 0BT80ZZ, 0BT84ZZ, 0BT90ZZ, 0BT94ZZ, 0BTB0ZZ, 0BTB4ZZ, 0BBC4ZZ, 0BBD4ZZ, 0BBF4ZZ, 0BBG4ZZ, 0BBH4ZZ, 0BBJ4ZZ, 0BBK4ZZ, 0BBL4ZZ, 0BQK0ZZ, 0BQK3ZZ, 0BQK4ZZ, 0BQK7ZZ, 0BQK8ZZ, 0BQL0ZZ, 0BQL3ZZ, 0BQL4ZZ, 0BQL7ZZ, 0BQL8ZZ, 0BQM0ZZ, 0BQM3ZZ, 0BQM4ZZ, 0BQM7ZZ, 0BQM8ZZ, 0B5K0ZZ, 0B5K3ZZ, 0B5K7ZZ, 0B5L0ZZ, 0B5L3ZZ, 0B5L7ZZ, 0B5M0ZZ, 0B5M3ZZ, 0B5M7ZZ, 0BBK0ZZ, 0BBK3ZZ, 0BBK7ZZ, 0BBL0ZZ, 0BBL3ZZ, 0BBL7ZZ, 0BBM0ZZ, 0BBM3ZZ, 0BBM7ZZ, 0B5C0ZZ, 0B5D0ZZ, 0B5F0ZZ, 0B5G0ZZ, 0B5H0ZZ, 0B5J0ZZ, 0B5K0ZZ, 0B5L0ZZ, 0B5M0ZZ, 0B5C3ZZ, 0B5D3ZZ, 0B5F3ZZ, 0B5G3ZZ, 0B5H3ZZ, 0B5J3ZZ, 0B5K3ZZ, 0B5L3ZZ, 0B5M3ZZ, 0B5C4ZZ, 0B5D4ZZ, 0B5F4ZZ, 0B5G4ZZ, 0B5H4ZZ, 0B5J4ZZ, 0B5K4ZZ, 0B5L4ZZ, 0B5M4ZZ, 0B5C7ZZ, 0B5C8ZZ, 0B5D7ZZ, 0B5D8ZZ, 0B5F7ZZ, 0B5F8ZZ, 0B5G7ZZ, 0B5G8ZZ, 0B5H7ZZ, 0B5H8ZZ, 0B5J7ZZ, 0B5J8ZZ, 0B5K7ZZ, 0B5K8ZZ, 0B5L7ZZ, 0B5L8ZZ, 0B5M7ZZ, 0B5M8ZZ, 0B538ZZ, 0B548ZZ, 0B558ZZ, 0B568ZZ, 0B578ZZ, 0B588ZZ, 0B598ZZ, 0B5B8ZZ, 0B5C8ZZ, 0B5D8ZZ, 0B5F8ZZ, 0B5G8ZZ, 0B5H8ZZ, 0B5J8ZZ, 0B5K8ZZ, 0B5L8ZZ, 0B5M8ZZ, 0BBC8ZZ, 0BBD8ZZ, 0BBF8ZZ, 0BBG8ZZ, 0BBH8ZZ, 0BBJ8ZZ, 0BBK8ZZ, 0BBL8ZZ, 0BBM4ZZ, 0BBM8ZZ, 0B5C0ZZ, 0B5C3ZZ, 0B5C7ZZ, 0B5D0ZZ, 0B5D3ZZ, 0B5D7ZZ, 0B5F0ZZ, 0B5F3ZZ, 0B5F7ZZ, 0B5G0ZZ, 0B5G3ZZ, 0B5G7ZZ, 0B5H0ZZ, 0B5H3ZZ, 0B5H7ZZ, 0B5J0ZZ, 0B5J3ZZ, 0B5J7ZZ, 0B5K0ZZ, 0B5K3ZZ, 0B5K7ZZ, 0B5L0ZZ, 0B5L3ZZ, 0B5L7ZZ, 0B5M0ZZ, 0B5M3ZZ, 0B5M7ZZ, 0BBC0ZZ, 0BBC3ZZ, 0BBC7ZZ, 0BBD0ZZ, 0BBD3ZZ, 0BBD7ZZ, 0BBF0ZZ, 0BBF3ZZ, 0BBF7ZZ, 0BBG0ZZ, 0BBG3ZZ, 0BBG7ZZ, 0BBH0ZZ, 0BBH3ZZ, 0BBH7ZZ, 0BBJ0ZZ, 0BBJ3ZZ, 0BBJ7ZZ, 0BBK0ZZ, 0BBK3ZZ, 0BBK7ZZ, 0BBL0ZZ, 0BBL3ZZ, 0BBL7ZZ, 0BBM0ZZ, 0BBM3ZZ, 0BBM7ZZ, 0BBC4ZZ, 0BBD4ZZ, 0BBF4ZZ, 0BBG4ZZ, 0BBH4ZZ, 0BBJ4ZZ, 0BBK4ZZ, 0BBL4ZZ, 0BTH4ZZ, 0BBK0ZZ, 0BBK3ZZ, 0BBK7ZZ, 0BBL0ZZ, 0BBL3ZZ, 0BBL7ZZ, 0BTH0ZZ, 0BTC4ZZ, 0BTD4ZZ, 0BTF4ZZ, 0BTG4ZZ, 0BTJ4ZZ, 0BTC0ZZ, 0BTD0ZZ, 0BTF0ZZ, 0BTG0ZZ, 0BTJ0ZZ, 0BTK4ZZ, 0BTL4ZZ, 0BTM4ZZ, 01B30ZZ, 01BL0ZZ, 0BTK0ZZ, 0BTL0ZZ, 0BTM0ZZ, 0PB10ZZ, 0PB20ZZ, 0B5K0ZZ, 0B5K3ZZ, 0B5K7ZZ, 0B5L0ZZ, 0B5L3ZZ, 0B5L7ZZ, 0B5M0ZZ, 0B5M3ZZ, 0B5M7ZZ, 0BBM0ZZ, 0BBM3ZZ, 0BBM7ZZ	

(*Continued*)

Table 3 (Continued).

Lung Resection Pre- or Post-Index		**Algorithm Points if ≥ 1 Code Present**
Lung Biopsy Pre- or Post-index		+4
CPT	31620, 31622, 31623, 31624, 31625, 31628, 31629, 31632, 31633, 31652, 31653, 31654, 31717, 32095, 32096, 32097, 32098, 32400, 32402, 32405, 32602, 32606, 32607, 32608, 32609, 38753, 39000, 39010, 39400, 39401, 39402	
ICD-9 procedures	33.20, 33.24, 33.25, 33.26, 33.27, 33.28, 34.20, 34.23, 34.24, 34.25, 34.26, 34.27	
ICD-10 procedures	0B930ZX, 0B933ZX, 0B934ZX, 0B937ZX, 0B938ZX, 0B940ZX, 0B943ZX, 0B944ZX, 0B947ZX, 0B948ZX, 0B950ZX, 0B953ZX, 0B954ZX, 0B957ZX, 0B958ZX, 0B960ZX, 0B963ZX, 0B964ZX, 0B967ZX, 0B968ZX, 0B970ZX, 0B973ZX, 0B974ZX, 0B977ZX, 0B978ZX, 0B980ZX, 0B983ZX, 0B984ZX, 0B987ZX, 0B988ZX, 0B990ZX, 0B993ZX, 0B994ZX, 0B997ZX, 0B998ZX, 0B9B0ZX, 0B9B3ZX, 0B9B4ZX, 0B9B7ZX, 0B9B8ZX, 0B9C0ZX, 0B9C3ZX, 0B9C4ZX, 0B9C7ZX, 0B9C8ZX, 0B9D0ZX, 0B9D3ZX, 0B9D4ZX, 0B9D7ZX, 0B9D8ZX, 0B9F0ZX, 0B9F3ZX, 0B9F4ZX, 0B9F7ZX, 0B9F8ZX, 0B9G0ZX, 0B9G3ZX, 0B9G4ZX, 0B9G7ZX, 0B9G8ZX, 0B9H0ZX, 0B9H3ZX, 0B9H4ZX, 0B9H7ZX, 0B9H8ZX, 0B9J0ZX, 0B9J3ZX, 0B9J4ZX, 0B9J7ZX, 0B9J8ZX, 0B9K0ZX, 0B9K3ZX, 0B9K4ZX, 0B9K7ZX, 0B9K8ZX, 0B9L0ZX, 0B9L3ZX, 0B9L4ZX, 0B9L7ZX, 0B9L8ZX, 0B9M0ZX, 0B9M3ZX, 0B9M4ZX, 0B9M7ZX, 0B9M8ZX, 0B9N0ZX, 0B9N3ZX, 0B9N4ZX, 0B9N8ZX, 0B9P0ZX, 0B9P3ZX, 0B9P4ZX, 0B9P8ZX, 0B9R0ZX, 0B9R3ZX, 0B9R4ZX, 0B9S0ZX, 0B9S3ZX, 0B9S4ZX, 0B9T0ZX, 0B9T3ZX, 0B9T4ZX, 0BB30ZX, 0BB33ZX, 0BB34ZX, 0BB37ZX, 0BB38ZX, 0BB40ZX, 0BB43ZX, 0BB44ZX, 0BB47ZX, 0BB48ZX, 0BB50ZX, 0BB53ZX, 0BB54ZX, 0BB57ZX, 0BB58ZX, 0BB60ZX, 0BB63ZX, 0BB64ZX, 0BB67ZX, 0BB68ZX, 0BB70ZX, 0BB73ZX, 0BB74ZX, 0BB77ZX, 0BB78ZX, 0BB80ZX, 0BB83ZX, 0BB84ZX, 0BB87ZX, 0BB88ZX, 0BB90ZX, 0BB93ZX, 0BB94ZX, 0BB97ZX, 0BB98ZX, 0BBB0ZX, 0BBB3ZX, 0BBB4ZX, 0BBB7ZX, 0BBB8ZX, 0BBC0ZX, 0BBC3ZX, 0BBC4ZX, 0BBC7ZX, 0BBC8ZX, 0BBD0ZX, 0BBD3ZX, 0BBD4ZX, 0BBD7ZX, 0BBD8ZX, 0BBF0ZX, 0BBF3ZX, 0BBF4ZX, 0BBF7ZX, 0BBF8ZX, 0BBG0ZX, 0BBG3ZX, 0BBG4ZX, 0BBG7ZX, 0BBG8ZX, 0BBH0ZX, 0BBH3ZX, 0BBH4ZX, 0BBH7ZX, 0BBH8ZX, 0BBJ0ZX, 0BBJ3ZX, 0BBJ4ZX, 0BBJ7ZX, 0BBJ8ZX, 0BBK0ZX, 0BBK3ZX, 0BBK4ZX, 0BBK7ZX, 0BBK8ZX, 0BBL0ZX, 0BBL3ZX, 0BBL4ZX, 0BBL7ZX, 0BBL8ZX, 0BBM0ZX, 0BBM3ZX, 0BBM4ZX, 0BBM7ZX, 0BBM8ZX, 0BBN0ZX, 0BBN3ZX, 0BBN4ZX, 0BBN8ZX, 0BBP0ZX, 0BBP3ZX, 0BBP4ZX, 0BBP8ZX, 0BBR0ZX, 0BBR3ZX, 0BBR4ZX, 0BBS0ZX, 0BBS3ZX, 0BBS4ZX, 0BBT0ZX, 0BBT3ZX, 0BBT4ZX, 0BD34ZX, 0BD38ZX, 0BD44ZX, 0BD48ZX, 0BD54ZX, 0BD58ZX, 0BD64ZX, 0BD68ZX, 0BD74ZX, 0BD78ZX, 0BD84ZX, 0BD88ZX, 0BD94ZX, 0BD98ZX, 0BDB4ZX, 0BDB8ZX, 0BDC4ZX, 0BDC8ZX, 0BDD4ZX, 0BDD8ZX, 0BDF4ZX, 0BDF8ZX, 0BDG4ZX, 0BDG8ZX, 0BDH4ZX, 0BDH8ZX, 0BDJ4ZX, 0BDJ8ZX, 0BDK4ZX, 0BDK8ZX, 0BDL4ZX, 0BDL8ZX, 0BDM4ZX, 0BDM8ZX, 0BDN0ZX, 0BDN3ZX, 0BDN4ZX, 0BDP0ZX, 0BDP3ZX, 0BDP4ZX, 0W980ZX, 0W983ZX, 0W984ZX, 0W990ZX, 0W993ZX, 0W994ZX, 0W9B0ZX, 0W9B3ZX, 0W9B4ZX, 0W9C0ZX, 0W9C3ZX, 0W9C4ZX, 0WB80ZX, 0WB83ZX, 0WB84ZX, 0WB8XZX, 0WBC0ZX, 0WBC3ZX, 0WBC4ZX	
≥10 days with a Lung Cancer Diagnosis Code Pre- or Post-index		+3
ICD-9-CM dx	162.2, 162.3, 162.4, 162.5, 162.8, 162.9	
ICD-10-CM dx	C34.00, C34.01, C34.02, C34.10, C34.11, C34.12, C34.2, C34.30, C34.31, C34.32, C34.80, C34.81, C34.82, C34.90, C34.91, C34.92	

(*Continued*)

Table 3 (Continued).

Lung Resection Pre- or Post-Index		**Algorithm Points if ≥ 1 Code Present**
PET Scan Post-index		+2
HCPC	G0125, G0126, G0210, G0211, G0212, G0234, G0235	
CPT	78112, 78113, 78114, 78115, 78116, 78811	
ICD-9-PCS	92.15	
ICD-10-PCS	CB32KZZ, CB32YZZ, CB3YYZZ	
ICD-9-CM Lung Cancer Diagnosis Code on an Outpatient Claim Pre- or Post-index		+1
ICD-9-CM dx	162.2, 162.3, 162.4, 162.5, 162.8, 162.9	
ICD-10-CM dx	C34.00, C34.01, C34.02, C34.10, C34.11, C34.12, C34.2, C34.30, C34.31, C34.32, C34.80, C34.81, C34.82, C34.90, C34.91, C34.92	
Pleurisy, Pleural Effusion, or Empyema Post-index		+1
ICD-9-CM dx	510.0, 510.9, 511.0, 511.1, 511.81, 511.89, 511.9	
ICD-10-CM dx	J86.0, J86.9, J90, J91.0, J91.8, J92.0, J92.9, J94.0, J94.1, J94.2, J94.8, J94.9, R09.1	
Male (+1 point)		+1
Chronic Pulmonary Disease (CCI) Pre-index		+1
ICD-9-CM dx	416.8, 416.9, 490, 491.0, 491.1, 491.20, 491.21, 491.22, 491.8, 491.9, 492.0, 492.8, 493.00, 493.01, 493.02, 493.10, 493.11, 493.12, 493.20, 493.21, 493.22, 493.81, 493.82, 493.90, 493.91, 493.92, 494.0, 494.1, 495.0, 495.1, 495.2, 495.3, 495.4, 495.5, 495.6, 495.7, 495.8, 495.9, 496, 500, 501, 502, 503, 504, 505, 506.4, 508.1, 508.8	
ICD-10-CM dx	I27.2, I27.20, I27.21, I27.22, I27.23, I27.24, I27.29, I27.81, I27.82, I27.83, I27.89, I27.9, J40, J41.0, J41.1, J41.8, J42, J43.0, J43.1, J43.2, J43.8, J43.9, J44.0, J44.1, J44.9, J45.20, J45.21, J45.22, J45.30, J45.31, J45.32, J45.40, J45.41, J45.42, J45.50, J45.51, J45.52, J45.901, J45.902, J45.909, J45.990, J45.991, J45.998, J47.0, J47.1, J47.9, J60, J61, J62.0, J62.8, J63.0, J63.1, J63.2, J63.3, J63.4, J63.5, J63.6, J64, J65, J66.0, J66.1, J66.2, J66.8, J67.0, J67.1, J67.2, J67.3, J67.4, J67.5, J67.6, J67.7, J67.8, J67.9, J68.4, J70.1, J70.3, J82.83, J84.170, J84.178	
Enlargement of Lymph Nodes Post-index		−1
ICD-9-CM dx	785.6	
ICD-10-CM dx	R59.0, R59.1, R59.9	
Malignant Neoplasm Of Main Bronchus Post-index		−1
ICD-9-CM	162.2	
ICD-10-CM	C34.00, C34.01, C34.02	

Notes: Index is the first ICD-9-CM code of 162.2 through 162.9 in the Medicare claims. The pre-index period used in our study was one year. Patients were followed until the earlier of death or one-year post index. Machine read-able code lists are available at https://doi.org/10.5281/zenodo.5095308. [a] Positive and negative points are summed for each patient. A score of ≥5 points denotes the patient is an incident (newly diagnosed) NSCLC case at the first instance of ICD-9-CM code 162.2–162.9 in the claims dataset. [b]ICD-10-CM and ICD-10-PCS codes recommended based on Optum360 Encoder Pro, Centers for Medicare & Medicaid Services general equivalence mappings, and coding and clinical expert review.

Abbreviations: CPT, current procedural terminology; dx, diagnosis; CCI, Charlson Comorbidity Index; HCPC, Healthcare Common Procedure Code; ICD-9-CM, International Classification of Diseases, Ninth Revision, Clinical Modification; ICD-10-CM, International Classification of Diseases, 10th Revision, Clinical Modification; PCS, procedure coding system.

Table 4 Sensitivity and Specificity of Point-Based Algorithm by Selected Characteristics from the Surveillance, Epidemiology, and End Results Registry: Univariate Logistic Regression Analysis

Variable	Sensitivity			Specificity		
	n	Odds Ratio (95% CI)	FDR p-value	n	Odds Ratio (95% CI)	FDR p-value
Current Reason for Entitlement to Medicare at Index Year						
Old age and survivors insurance (age)	52,745	1.81 (0.76, 4.27)	0.2387	179,076	0.13 (0.05, 0.35)	0.0002
Disability insurance benefits	60	0.79 (0.24, 2.63)	0.7606	2929	13.94 (3.48, 55.77)	0.0004
End-stage renal disease	71	0.4 (0.11, 1.41)	0.2107	173	0.79 (0.19, 3.17)	0.7896
Follow-up time:						
<3 months	13,733	0.32 (0.3, 0.35)	<0.0001	17,443	0.4 (0.35, 0.46)	<0.0001
3–5.99 months	6920	1.41 (1.28, 1.56)	<0.0001	7781	0.21 (0.18, 0.24)	<0.0001
6–8.99 months	4582	1.87 (1.67, 2.1)	<0.0001	6325	0.14 (0.13, 0.17)	<0.0001
9–11.9 months	3467	1.95 (1.71, 2.21)	<0.0001	5201	0.17 (0.14, 0.19)	<0.0001
1–1.99 years	10,444	1.33 (1.23, 1.44)	<0.0001	36,141	0.74 (0.66, 0.83)	<0.0001
2–2.99 years	5134	1.19 (1.07, 1.33)	0.0025	27,996	2.66 (2.19, 3.23)	<0.0001
3–3.99 years	3220	1.07 (0.94, 1.22)	0.4061	22,546	5.34 (3.96, 7.21)	<0.0001
4–4.99 years	2104	1.15 (0.98, 1.36)	0.1224	18,864	9.32 (6.06, 14.34)	<0.0001
5–5.99 years	1494	0.91 (0.75, 1.1)	0.4026	15,660	9.39 (5.82, 15.14)	<0.0001
6–6.99 years	1066	0.9 (0.72, 1.13)	0.4461	13,260	16.75 (8.36, 33.54)	<0.0001
>7 years	714	0.95 (0.73, 1.25)	0.7896	11,031	12.23 (6.35, 23.55)	<0.0001
Medicare Status Code at Index						
Aged	52,477	1.11 (0.72, 1.7)	0.7064	180,754	0.6 (0.3, 1.21)	0.2111
Aged with end-stage renal disease	340	0.95 (0.6, 1.5)	0.8585	401	0.62 (0.25, 1.49)	0.3661
Disabled	55	0.68 (0.18, 2.62)	0.6468	1036	3.22 (1.04, 10)	0.0675
Numbers of Primary Cancers (Any Cancer Type)						
1	38,347	1.5 (1.4, 1.62)	<0.0001	14,979	1.02 (0.89, 1.17)	0.832
2	11,545	0.74 (0.68, 0.8)	<0.0001	3202	0.98 (0.85, 1.14)	0.8585
≥3	2986	0.59 (0.51, 0.68)	<0.0001	658	0.98 (0.73, 1.31)	0.9067
Race						
White	45,442	0.9 (0.82, 0.98)	0.0306	152,917	0.62 (0.53, 0.73)	<0.0001
Black	4411	1.15 (1.03, 1.29)	0.0247	13,939	1.18 (0.97, 1.44)	0.1418
Other	839	0.77 (0.6, 1)	0.0794	4028	2.19 (1.36, 3.54)	0.0027
Asian	1478	1.1 (0.91, 1.34)	0.4015	6671	1.66 (1.19, 2.3)	0.0049
Hispanic	538	1.35 (0.99, 1.85)	0.0822	3842	2.96 (1.68, 5.22)	0.0004
North American Native	126	1.22 (0.65, 2.31)	0.6079	603	1.79 (0.57, 5.56)	0.3943
Sex						
Male	28,050	0.9 (0.84, 0.96)	0.0025	75,841	0.76 (0.69, 0.84)	<0.0001
Female	24,828	1.12 (1.04, 1.19)	0.0025	106,407	1.31 (1.19, 1.45)	<0.0001
Age Group at Index						
66–69	9900	1.34 (1.23, 1.45)	<0.0001	35,625	0.63 (0.56, 0.7)	<0.0001
70–74	13,610	1.25 (1.16, 1.34)	<0.0001	40,330	0.64 (0.57, 0.71)	<0.0001
75–79	13,413	1.05 (0.98, 1.14)	0.232	37,003	0.81 (0.72, 0.91)	0.0009
≥80	15,955	0.6 (0.56, 0.65)	<0.0001	69,290	2.91 (2.55, 3.31)	<0.0001
Charlson Comorbidities						
Myocardial infarction	6781	1.12 (1.01, 1.25)	0.044	13,397	0.46 (0.4, 0.53)	<0.0001
Congestive heart failure	14,895	1.05 (0.97, 1.13)	0.2688	38,678	0.68 (0.61, 0.76)	<0.0001
Peripheral vascular disease	18,059	1.07 (1, 1.15)	0.0741	41,694	0.53 (0.48, 0.59)	<0.0001
Dementia	3147	0.7 (0.61, 0.81)	<0.0001	19,466	3.45 (2.64, 4.5)	<0.0001
Cerebrovascular disease	16,486	1.09 (1.02, 1.18)	0.0292	39,621	0.5 (0.46, 0.56)	<0.0001
Chronic pulmonary disease	38,128	3.33 (3.08, 3.59)	<0.0001	55,685	0.06 (0.05, 0.07)	<0.0001
Rheumatic disease	3373	1.17 (1.02, 1.34)	0.044	9284	0.94 (0.76, 1.17)	0.6682
Peptic ulcer disease	2405	1.2 (1.02, 1.41)	0.0427	5472	0.68 (0.54, 0.87)	0.0042
Mild liver disease	7541	0.7 (0.64, 0.77)	<0.0001	11,989	0.32 (0.28, 0.36)	<0.0001
Hemiplegia or paraplegia	1620	0.71 (0.57, 0.87)	0.0021	4392	0.66 (0.51, 0.86)	0.0047

(*Continued*)

Table 4 (Continued).

Variable	Sensitivity			Specificity		
	n	Odds Ratio (95% CI)	FDR p-value	n	Odds Ratio (95% CI)	FDR p-value
Renal disease	8134	0.87 (0.79, 0.96)	0.0096	21,285	0.78 (0.68, 0.89)	0.0009
Diabetes without chronic complications	17,468	1.15 (1.07, 1.23)	0.0004	55,406	0.86 (0.78, 0.96)	0.011
Moderate or severe liver disease	321	0.85 (0.55, 1.29)	0.5169	889	0.69 (0.38, 1.24)	0.28
Diabetes with chronic complications	4966	1.09 (0.97, 1.22)	0.2193	17,023	1.02 (0.86, 1.21)	0.8557
AIDS/HIV	80	0.8 (0.32, 2.02)	0.7064	180	0.31 (0.13, 0.75)	0.017
Charlson Comorbidity Score						
0	3955	0.37 (0.33, 0.42)	<0.0001	47,407	13.77 (10.27, 18.47)	<0.0001
1	10,473	1.09 (1, 1.18)	0.0693	40,057	1.33 (1.17, 1.51)	<0.0001
2	10,962	1.19 (1.1, 1.29)	<0.0001	27,003	0.55 (0.49, 0.61)	<0.0001
3	8828	1.1 (1.01, 1.21)	0.0527	20,142	0.58 (0.51, 0.66)	<0.0001
4	6451	1.16 (1.04, 1.28)	0.0103	13,540	0.56 (0.48, 0.65)	<0.0001
5	4367	1.08 (0.95, 1.22)	0.3003	9709	0.56 (0.47, 0.67)	<0.0001
6	2921	1.2 (1.03, 1.4)	0.0314	6576	0.57 (0.46, 0.7)	<0.0001
7	1863	1.02 (0.84, 1.24)	0.8964	4296	0.64 (0.49, 0.84)	0.0023
8	1218	0.96 (0.76, 1.2)	0.7606	2682	0.56 (0.41, 0.77)	0.0009
9	673	1.24 (0.91, 1.7)	0.2322	1614	0.69 (0.44, 1.07)	0.1387
10	370	0.85 (0.54, 1.36)	0.5861	883	0.62 (0.35, 1.1)	0.1434
>10	220	0.77 (0.41, 1.44)	0.4961	707	0.43 (0.25, 0.72)	0.0033
AJCC Stage (versions may Vary Across Years)						
12 – IA	7066	0.56 (0.5, 0.62)	<0.0001	725	2.13 (1.45, 3.12)	0.0002
15 – IB	6007	1.38 (1.24, 1.54)	<0.0001	716	0.91 (0.68, 1.22)	0.5931
32 – IIA	583	2.3 (1.67, 3.16)	<0.0001	108	0.55 (0.27, 1.11)	0.1362
33 – IIB	2004	2.34 (1.97, 2.78)	<0.0001	225	0.58 (0.37, 0.91)	0.0306
52 – IIIA	4689	2.48 (2.21, 2.78)	<0.0001	1,305	0.26 (0.21, 0.31)	<0.0001
53 – IIIB	8,097	1.56 (1.43, 1.71)	<0.0001	2,614	0.4 (0.35, 0.46)	<0.0001
70 – IV	19,936	0.63 (0.59, 0.68)	<0.0001	9,198	1.11 (1, 1.24)	0.0799
88 – NA	463	0.14 (0.09, 0.21)	<0.0001	157	0.75 (0.39, 1.44)	0.4744
90 – OCCULT	817	0.91 (0.7, 1.18)	0.5574	319	1.25 (0.82, 1.9)	0.3833
99 – Unknown stage	3,216	0.77 (0.67, 0.88)	0.0002	3,472	4.36 (3.59, 5.31)	<0.0001
Evidence of a Primary Payer Other than Medicare						
No	48,106	2.69 (2.39, 3.04)	<0.0001	172,196	1.3 (1.07, 1.58)	0.0151
Yes	4,772	0.37 (0.33, 0.42)	<0.0001	10,052	0.77 (0.63, 0.93)	0.0151
State at Index						
California	14,897	0.9 (0.84, 0.97)	0.0069	55,558	1.59 (1.42, 1.8)	<0.0001
Connecticut	3303	1.01 (0.89, 1.16)	0.8587	11,213	1.13 (0.91, 1.39)	0.3627
Georgia	6328	1.19 (1.08, 1.32)	0.0011	20,926	0.71 (0.62, 0.82)	<0.0001
Hawaii	544	0.75 (0.55, 1.01)	0.0918	2435	1.85 (1.05, 3.26)	0.0555
Iowa	2997	0.63 (0.54, 0.73)	<0.0001	11,110	1.05 (0.85, 1.3)	0.7064
Kentucky	5133	1.03 (0.92, 1.16)	0.6895	13,928	0.61 (0.52, 0.71)	<0.0001
Louisiana	3421	1.24 (1.09, 1.41)	0.0025	11,062	0.71 (0.59, 0.85)	0.0004
Michigan	3908	1.18 (1.04, 1.33)	0.0216	12,213	0.86 (0.72, 1.04)	0.1664
New Jersey	7771	1.3 (1.18, 1.43)	<0.0001	24,684	0.93 (0.81, 1.07)	0.3787
New Mexico	931	0.84 (0.65, 1.09)	0.2456	4530	1.45 (1, 2.11)	0.0799
Utah	598	0.97 (0.74, 1.28)	0.8587	3999	2.06 (1.29, 3.28)	0.0046
Washington	3047	0.61 (0.53, 0.71)	<0.0001	10,590	0.94 (0.76, 1.15)	0.6079
Tumor Histology at Index						
NSCLC and SCLC	86	2.45 (1.15, 5.24)	0.0345	–	n/a	-
NSCLC and other	124	1.42 (0.74, 2.73)	0.3757	–	n/a	-
NSCLC only	52,668	0.56 (0.34, 0.91)	0.0327	–	n/a	-
Other and NSCLC	–	n/a	-	35	0.93 (0.2, 4.37)	0.9383
Other only	–	n/a	–	10,422	19.85 (17.01, 23.16)	<0.0001

(*Continued*)

Table 4 (Continued).

Variable	Sensitivity			Specificity		
	n	Odds Ratio (95% CI)	FDR p-value	n	Odds Ratio (95% CI)	FDR p-value
SCLC and NSCLC	–	n/a	–	49	0.19 (0.08, 0.46)	0.0004
SCLC only	–	n/a	–	8327	0.05 (0.05, 0.06)	<0.0001
Tumor histology	–	–	–	–	–	–
Lung cancer–other	–	–	–	10,460	19.67 (16.87, 22.94)	<0.0001
SCLC	–	–	–	8379	0.05 (0.04, 0.06)	<0.0001
Laterality						
Not a paired site	35	3.17 (1.19, 8.45)	0.035	25	0.12 (0.04, 0.31)	<0.0001
Right: origin of primary	29,147	1.14 (1.07, 1.22)	0.0002	9116	0.65 (0.58, 0.72)	<0.0001
Left: origin of primary	21,456	1.16 (1.08, 1.24)	<0.0001	6703	0.68 (0.6, 0.76)	<0.0001
Only one side involved, right or left origin unspecified	169	0.3 (0.16, 0.56)	0.0002	260	4.37 (2.14, 8.92)	0.0002
Bilateral involvement, lateral origin unknown; stated to be single primary	647	0.49 (0.36, 0.67)	<0.0001	322	3.43 (1.86, 6.33)	0.0002
Paired site, but no information concerning laterality; midline tumor	1424	0.17 (0.13, 0.22)	<0.0001	2413	7.3 (5.53, 9.64)	<0.0001
Marital Status at Diagnosis						
Single (never married)	3948	1.02 (0.9, 1.15)	0.8216	1447	1.09 (0.88, 1.35)	0.495
Married (including common law)	26,859	1.11 (1.04, 1.18)	0.0051	7789	0.56 (0.5, 0.62)	<0.0001
Separated	330	0.88 (0.6, 1.3)	0.5931	115	1.28 (0.6, 2.72)	0.6009
Divorced	4868	0.94 (0.84, 1.05)	0.3833	1928	1.15 (0.95, 1.38)	0.203
Widowed	14,959	0.95 (0.88, 1.02)	0.1925	6791	1.74 (1.54, 1.96)	<0.0001
Unknown	1908	0.76 (0.63, 0.91)	0.0056	765	1.11 (0.84, 1.47)	0.5551
Evidence of a Primary Payer Other than Medicare						
Insurance, NOS	558	0.67 (0.48, 0.93)	0.028	204	0.94 (0.55, 1.6)	0.8585
Private insurance: managed care, HMO, or PPO	1003	0.39 (0.3, 0.52)	<0.0001	243	0.78 (0.5, 1.21)	0.3415
Medicaid	181	0.61 (0.33, 1.13)	0.1677	52	2.83 (0.67, 11.95)	0.2152
Medicaid – administered through a managed care plan	55	3.78 (0.98, 14.64)	0.0806	13	0.9 (0.1, 8.04)	0.9383
Medicare/Medicare, NOS	9865	1.09 (1, 1.19)	0.0728	3714	1.04 (0.9, 1.2)	0.6453
Medicare with supplement, NOS	14,966	1.22 (1.12, 1.31)	<0.0001	4349	0.59 (0.51, 0.67)	<0.0001
Medicare – administered through a managed care plan	2085	0.91 (0.76, 1.09)	0.3691	605	1.01 (0.73, 1.4)	0.9672
Medicare with private supplement	9387	1.29 (1.17, 1.41)	<0.0001	2679	0.57 (0.49, 0.67)	<0.0001
Medicare with Medicaid eligibility	3981	1.4 (1.24, 1.59)	<0.0001	1627	0.97 (0.79, 1.2)	0.832
TRICARE	285	0.24 (0.15, 0.39)	<0.0001	92	1.08 (0.52, 2.21)	0.8652
Insurance status unknown	991	0.41 (0.32, 0.54)	<0.0001	1768	8.42 (5.92, 11.97)	<0.0001
SEER Primary Site						
C34.0 Main bronchus	1785	4.2 (3.65, 4.85)	<0.0001	1314	0.23 (0.2, 0.27)	<0.0001
C34.1 Upper lobe, lung	26,846	1.11 (1.04, 1.19)	0.0036	7463	0.78 (0.7, 0.87)	<0.0001
C34.2 Middle lobe, lung (right lung only)	2211	1 (0.85, 1.18)	0.9709	663	0.66 (0.5, 0.88)	0.0083
C34.3 Lower lobe, lung	14,901	0.99 (0.92, 1.07)	0.8964	3590	1.01 (0.87, 1.17)	0.9488
C34.8 Overlapping lesion of lung	540	1.82 (1.36, 2.43)	0.0002	214	0.59 (0.37, 0.93)	0.0361
C34.9 Lung, NOS	6595	0.38 (0.34, 0.42)	<0.0001	5595	3.12 (2.72, 3.58)	<0.0001
Region at Index						
Midwest	6905	0.9 (0.81, 1)	0.0642	23,323	0.94 (0.81, 1.09)	0.4841
Northeast	11,074	1.22 (1.13, 1.33)	<0.0001	35,897	0.99 (0.87, 1.11)	0.8557
South	14,882	1.19 (1.11, 1.28)	<0.0001	45,916	0.61 (0.55, 0.68)	<0.0001
Midwest	20,017	0.79 (0.74, 0.85)	<0.0001	77,112	1.62 (1.46, 1.8)	<0.0001
Urban/Rural						
Big metro (urban = 00 or 01)	28,181	0.91 (0.85, 0.97)	0.0083	95,499	1.25 (1.14, 1.38)	<0.0001
Metro (urban = 02 or 03)	15,536	1.09 (1.02, 1.18)	0.0221	54,248	0.98 (0.88, 1.09)	0.7472
Urban (urban = 04 or 05)	3133	0.94 (0.81, 1.08)	0.4326	11,215	0.88 (0.72, 1.07)	0.2642
Less urban (urban = 06 or 07)	4876	1.06 (0.94, 1.19)	0.4061	17,259	0.7 (0.6, 0.81)	<0.0001
Rural (urban = 08 or 09)	1142	1.18 (0.95, 1.47)	0.2023	4018	0.7 (0.52, 0.93)	0.0247

(*Continued*)

Table 4 (Continued).

Variable	Sensitivity			Specificity		
	n	Odds Ratio (95% CI)	FDR p-value	n	Odds Ratio (95% CI)	FDR p-value
Year of Diagnosis (Based on SEER)						
2005	8022	1.22 (1.12, 1.33)	<0.0001	2993	0.95 (0.82, 1.1)	0.5764
2006	7935	1.12 (1.02, 1.22)	0.0232	2923	0.82 (0.71, 0.95)	0.012
2007	7764	1.07 (0.97, 1.17)	0.2152	2746	0.86 (0.74, 1)	0.0804
2008	7536	1.12 (1.02, 1.23)	0.034	2722	1.12 (0.96, 1.32)	0.2076
2009	7450	0.98 (0.89, 1.08)	0.7342	2669	1.05 (0.89, 1.23)	0.6234
2010	7120	0.79 (0.72, 0.88)	<0.0001	2541	1.19 (1, 1.41)	0.0716
2011	7051	0.69 (0.62, 0.76)	<0.0001	2245	1.17 (0.98, 1.4)	0.1178
SEER Summary Stage						
Localized only	11,922	0.73 (0.68, 0.8)	<0.0001	1652	1.24 (1, 1.53)	0.0724
Regional by direct extension only	12,899	2.64 (2.44, 2.85)	<0.0001	3070	0.32 (0.28, 0.36)	<0.0001
Distant site(s)/node(s) involved	26,465	0.64 (0.6, 0.68)	<0.0001	11,463	0.94 (0.84, 1.05)	0.3711
Unknown/unstaged/unspecified	1592	0.53 (0.43, 0.64)	<0.0001	2654	5.97 (4.65, 7.67)	<0.0001

Abbreviations: AJCC, American Joint Committee on Cancer; CI, confidence interval; FDR, false discovery rate; HMO, health maintenance organization; n/a, odds ratio not estimable because all subjects were "non-lung cancer" based on the score; NA, not applicable (for variables only available in SEER for lung cancer cases); NOS, not otherwise specified; NSCLC, non-small cell lung cancer; PPO, preferred provider organization; SEER, Surveillance, Epidemiology, and End Results; SCLC, small cell lung cancer.

Table 5 Sensitivity and Specificity of Point-Based Algorithm by Selected Characteristics from the Surveillance, Epidemiology, and End Results Registry: Multivariable Logistic Regression Analysis

Variables	Odds Ratio (95% CI)	p-value	NNT
Associated with Increased or Decreased Sensitivity:			
SEER primary site Bronchus: Main	3.30 (2.76–3.94)	<0.0001	
Chronic pulmonary disease	3.23 (2.94–3.55)	<0.0001	
No evidence of a primary payer other than Medicare	2.19 (1.89–2.52)	<0.0001	
SEER Summary Stage 2000 Regional by direct extension only	1.83 (1.65–2.02)	<0.0001	
Numbers of primary cancers (any cancer type)=1	1.69 (1.54–1.86)	<0.0001	
State at index New Jersey	1.38 (1.23–1.54)	<0.0001	
Follow-up time 6–8.9 months	1.37 (1.18–1.59)	<0.0001	
AJCC stage IIIB	1.27 (1.14–1.43)	<0.0001	
Year of diagnosis (based on SEER) 2011	0.70 (0.62–0.79)	<0.0001	11.7
Age group at index ≥ 80	0.64 (0.58–0.70)	<0.0001	8.5
SEER primary site Lung NOS	0.63 (0.55–0.72)	<0.0001	4.9
AJCC stage IA	0.52 (0.45–0.60)	<0.0001	7.7
No information concerning laterality	0.34 (0.25–0.46)	<0.0001	3.4
Follow-up time <3 months	0.30 (0.27–0.34)	<0.0001	4.1
AJCC stage Not applicable	0.09 (0.05–0.15)	<0.0001	3.2
Associated with Increased or Decreased Specificity:			
Tumor histology at index: Other Lung Cancer only	14.75 (12.12–17.945)	<0.0001	
Follow-up time <3 months	6.31 (5.24–7.60)	<0.0001	
Age group at index ≥ 80	3.07 (2.16–4.37)	<0.0001	
No information concerning laterality	1.77 (1.46–2.14)	<0.0001	
SEER Summary Stage 2000 Regional by direct extension only	0.51 (0.41–0.62)	<0.0001	4.7
SEER primary site Bronchus: Main	0.29 (0.23–0.38)	<0.0001	3.4
Chronic pulmonary disease	0.22 (0.18–0.28)	<0.0001	5.9

Abbreviations: NNT, the number needed in the subgroup to result in one more false negative (sensitivity) or one more false positive (specificity); AJCC, American Joint Committee on Cancer; SEER, Surveillance, Epidemiology, and End Results.

the neural network algorithm, the sensitivity, PPV, and F-score for the point-based algorithm were similar in magnitude (difference in sensitivity = +0.72%, PPV = −4.71%, F-score −2.27%).

Algorithm performance was assessed using the F-score, a composite measure of a model's precision (PPV) and sensitivity, and we sought to optimize both. Sensitivity and PPV of our final algorithm approached or exceeded levels generally considered acceptable (≥70%),[23,24] although this is not an immutable threshold. Depending on the context for the analysis, the thresholds could vary.

Better algorithm performance was observed for all methods when applied after the initial lung cancer algorithm[15] (Approach B) vs to the entire sample (Approach A). NSCLC cases were easier to identify from among the patients with lung cancer as selected by the lung cancer algorithm, rather than searching for NSCLC cases without that prior information.

One advantage of the point-score algorithm is that only data from claims are needed to implement it. Some other algorithm validation studies use the reference standard data source (eg, registry or electronic medical record data) as a first algorithm step for identifying cases; this cannot be an algorithm step in real-world practice, however, since other researchers likely do not have access to the registry or electronic medical record reference standard (the reason they are using an algorithm). Our two-step algorithm (Approach B) utilizes only claims data for identifying the patient with NSCLC.

In this study, it was necessary to consider misclassification and bias in both the lung cancer and NSCLC point-score algorithms. The patient characteristics associated with statistically significant reduced sensitivity in the lung cancer algorithm[15] (patients ≥80 years, follow-up time in Medicare <3 months, and missing SEER data on stage, laterality, or site) were also observed in the NSCLC algorithm. Some additional characteristics associated with misclassification in the NSCLC algorithm were related to reduced algorithm specificity for site (SEER primary site of bronchus: main), stage (SEER summary stage 2000 of regional by direct extension only), and pre-index chronic pulmonary disease.

We previously described the external generalizability of the lung cancer algorithm.[15]

A strength of the NSCLC point-score algorithm is face validity of components. The heavy weighted algorithm components (≥2 points) are lung resection, biopsy, unique lung cancer days, and positron emission tomography (PET) scan. These procedures are more likely to be associated with NSCLC than SCLC, based on treatment guidelines.[17] For example, when SCLC is disseminated (in most patients at presentation), these patients are often not candidates for resection, may be more frequently biopsied at sites other than lung, and may not have a PET/computed tomography scan performed. These algorithm components and concepts remain relevant, although future work could include evaluating the concepts and/or weights in a more recent dataset (eg, if NSCLC or non-NSCLC cancer stage distribution or diagnostic procedures change).

This study was limited by generalizability to patients in health maintenance organization plans and commercial claims data sources with individuals older than 65 who were not included in this study, as described previously.[15] The algorithms in this study relied on ICD-9 diagnoses and procedures. Although more recent US real-world data contain diagnoses coded in ICD-10, these were not yet effectively available in US real-world data at the time of our study. We provided proposed crosswalks to convert ICD-9 to ICD-10 codes for the single-score system (Table 3). The final point-based algorithm criteria did not include personal history of tobacco use (ICD-9-CM V15.82) which is not robustly coded in US administrative healthcare data (eg, if these diagnosis codes are not associated with a reimbursable health procedure to generate a medical claim). Additional information on smoking history (eg, smoking pack-years) may contribute meaningful information to differentiate NSCLC from other types of lung cancer but was not available in this data source. The algorithm may perform differently in administrative healthcare data sources outside the US, as different coding adaptations of ICD and different real-world reimbursement requirements/environment in other countries influence the performance of real-world data algorithms.[25]

Conclusion

Our study developed and validated a practical, 10-variable, point-based algorithm for identifying incident NSCLC cases in a US claims database based on a previously validated incident lung cancer algorithm. We developed the algorithm using diagnostic and procedural codes instead of medications to support the algorithm's longer-term relevance and reliability and validated the algorithm using a criterion validity approach. The fit for purpose of any RWE algorithm is dependent on its context of use[1] and both internal validity (eg, algorithm performance within this study) and external validity (eg, generalizability of the algorithm to the real-world data source where it will be applied for conducting future

research) should be assessed by researchers considering whether to use the algorithm.[26] In our view, based on the performance of the final point-score in terms of its sensitivity and PPV, the final point-score NSCLC algorithm is likely fit for purpose for general disease state studies including burden of illness and treatment patterns. Researchers should consider the algorithm's performance in the context of their study question and data source, as described in the Certainty Framework for real-world data variables.[27] The implementation of a previously validated, broader lung cancer algorithm increased the performance of the point-based algorithm, which may be an important consideration for researchers building algorithms that comprise a subtype of the disease.

Data Sharing Statement

The dataset used for the current study is not publicly available due to SEER-Medicare Data Use Agreement restrictions. However, researchers may obtain access to SEER-Medicare data by submitting a proposal (details for submitting proposals are available at https://healthcaredelivery.cancer.gov/seermedicare/obtain/).

Ethics Statement

The protocol was reviewed and considered exempt by Quorum Review IRB prior to National Cancer Institute approval of the SEER-Medicare data for this study.

Acknowledgments

The authors thank Elaine Yanisko of IMS, Inc. for support in triaging data-related questions with staff at the National Cancer Institute, Shannon Gardell and Colleen Dumont of Evidera for their expert writing and editorial reviews of the manuscript, and Yushi Liu of Eli Lilly and Company for statistical peer review. For feasibility analysis and creating the analytic dataset, the authors thank Tim Ellington of Delisle Associates LTD. For validation of analytic programs, the authors thank Jessica Mitroi of Eli Lilly and Company. For quality review of the final manuscript, the authors thank Nancy Hedlund of MedNavigate LLC. .JB and DRN are joint senior authors for this study.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This study was funded by Eli Lilly and Company. Medical writing assistance was provided by Shannon Gardell of Evidera and was funded by Eli Lilly and Company. Evidera complied with international guidelines for Good Publication Practice (GPP3).

Disclosure

JB, DRN, KMS, and YJH are employees and shareholders of Eli Lilly and Company. YKL was an employee of Eli Lilly and Company during the conduct of the study. ALH is an employee of the University of Cincinnati and reports grants from Eli Lilly during the conduct of the study. The authors report no other conflicts of interest in this work.

References

1. Beyrer J, Abedtash H, Hornbuckle K, Murray JF. A review of stakeholder recommendations for defining fit-for-purpose real-world evidence algorithms. *J Comp Eff Res*. 2022;11(7):499–511. doi:10.2217/cer-2022-0006
2. National Cancer Institute [Internet]. Cancer Stat Facts: lung and Bronchus Cancer. Available from: https://seer.cancer.gov/statfacts/html/lungb.html. Accessed November 11, 2022.
3. United States Preventive Services Task Force [Internet]. Final recommendation statement: lung cancer: screening; 2021. Available from https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/lung-cancer-screening. Accessed November 11, 2022.

4. American Cancer Society [Internet]. What is lung cancer? Available from: https://www.cancer.org/cancer/lung-cancer/about/what-is.html. Accessed November. 10, 2022.
5. Uno H, Ritzwoller DP, Cronin AM, Carroll NM, Hornbrook MC, Hassett MJ. Determining the time of cancer recurrence using claims or electronic medical record data. *JCO Clin Cancer Inform*. 2018;2:1–10. doi:10.1200/CCI.17.00163
6. Abraha I, Montedori A, Serraino D, et al. Accuracy of administrative databases in detecting primary breast cancer diagnoses: a systematic review. *BMJ Open*. 2018;8(7):e019264. doi:10.1136/bmjopen-2017-019264
7. van Walraven C, Austin P. Administrative database research has unique characteristics that can risk biased results. *J Clin Epidemiol*. 2012;65 (2):126–131. doi:10.1016/j.jclinepi.2011.08.002
8. Chan AW, Fung K, Tran JM, et al. Application of recursive partitioning to derive and validate a claims-based algorithm for identifying keratinocyte carcinoma (nonmelanoma skin cancer). *JAMA Dermatol*. 2016;152(10):1122–1127. doi:10.1001/jamadermatol.2016.2609
9. Nordstrom BL, Simeone JC, Malley KG, et al. Validation of claims algorithms for progression to metastatic cancer in patients with breast, non-small cell lung, and colorectal cancer. *Front Oncol*. 2016;1(6):18.
10. Bergquist SL, Brooks GA, Keating NL, Landrum MB, Rose S. Classifying lung cancer severity with ensemble machine learning in health care claims data. *Proc Mach Learn Res*. 2017;68:25–38. doi:10.1016/j.csbj.2014.11.005
11. Turner RM, Chen YW, Fernandes AW. Validation of a case-finding algorithm for identifying patients with non-small cell lung cancer (NSCLC) in administrative claims databases. *Front Pharmacol*. 2017;30(8):883. doi:10.3389/fphar.2017.00883
12. Centers for Medicare and Medicaid Services [Internet]. Centers for Medicare and Medicaid Services. HCPCS - General Information. Baltimore (MD). Available from: https://www.cms.gov/Medicare/Coding/MedHCPCSGenInfo. Accessed August 18, 2022.
13. Nadpara P, Madhavan SS, Tworek C. Guideline-concordant timely lung cancer care and prognosis among elderly patients in the United States: a population-based study. *Cancer Epidemiol*. 2015;39(6):1136–1144. doi:10.1016/j.canep.2015.06.005
14. Wong ML, McMurry TL, Stukenborg GJ, et al. Impact of age and comorbidity on treatment of non-small cell lung cancer recurrence following complete resection: a nationally representative cohort study. *Lung Cancer*. 2016;102:108–117. doi:10.1016/j.lungcan.2016.11.002
15. Beyrer J, Nelson DR, Sheffield KM, Huang YJ, Ellington T, Hincapie AL. Development and validation of coding algorithms to identify patients with incident lung cancer in United States healthcare claims data. *Pharmacoepidemiol Drug Saf*. 2020;29(11):1465–1479. doi:10.1002/pds.5137
16. Nattinger AB, Laud PW, Bajorunaite R, Sparapani RA, Freeman JL. An algorithm for the use of Medicare claims data to identify women with incident breast cancer. *Health Serv Res*. 2004;39(6 Pt 1):1733–1749. doi:10.1111/j.1475-6773.2004.00315.x
17. National Comprehensive Cancer Network [Internet]. Plymouth Meeting (PA): national Comprehensive Cancer Network. National Comprehensive Cancer Network Guidelines. Available from: https://www.nccn.org/professionals/physician_gls/default.aspx. Accessed August. 18, 2022.
18. Zhao Z, Zhang R, Cox J, Duling D, Sarle W. Massively parallel feature selection: an approach based on variance preservation. *Mach Learn*. 2013;92(1):195–220. doi:10.1007/s10994-013-5373-4
19. Sullivan LM, Massaro JM, D'Agostino RB. D'Agostino RB Sr. Presentation of multivariate data for clinical use: the Framingham Study risk score functions. *Stat Med*. 2004;23(10):1631–1660. doi:10.1002/sim.1742
20. Zhang Z, Zhang H, Khanal MK. Development of scoring system for risk stratification in clinical medicine: a step-by-step tutorial. *Ann Transl Med*. 2017;5(21):436. doi:10.21037/atm.2017.08.22
21. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L, Explaining explanations: an overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). IEEE. 2018; 80–89.
22. Lipton ZC. The Mythos of Model Interpretability: in machine learning, the concept of interpretability is both important and slippery. *Queue*. 2018;16(3):31–57. doi:10.1145/3236386.3241340
23. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. 2nd ed. New York (NY): John Wiley and Sons; 2015:160–164.
24. Schumock GT, Lee TA, Pickard AS, et al. Mini-Sentinel methods — alternative methods for health outcomes of interest validation. FDA White Paper; 2013. Available from: https://www.sentinelinitiative.org/sites/default/files/surveillance-tools/validations-literature/Mini-Sentinel-Alternative-Methods-for-Health-Outcomes-of-Interest-Validation_0.pdf. Accessed August 18, 2022.
25. Schulman KL, Berenson K, Tina Shih YC, et al. A checklist for ascertaining study cohorts in oncology health services research using secondary data: report of the ISPOR oncology good outcomes research practices working group. *Value Health*. 2013;16(4):655–669. doi:10.1016/j.jval.2013.02.006
26. Singh S, Beyrer J, Zhou X, et al. Development and Evaluation of the Algorithm CErtaInty Tool (ACE-IT) to Assess Electronic Medical Record and Claims-based Algorithms' Fit for Purpose for Safety Outcomes. *Drug Saf*. 2022. doi:10.1007/s40264-022-01254-4
27. Cocoros NM, Arlett P, Dreyer NA, et al. The Certainty Framework for assessing real-world data in studies of medical product safety and effectiveness. *Clin Pharmacol Ther*. 2021;109(5):1189–1196. doi:10.1002/cpt.2045