RESEARCH

Sensitivity of Administrative Claims to Identify Incident Cases of Lung Cancer: A Comparison of 3 Health Plans

Scott D. Ramsey, MD, PhD; John F. Scoggins, PhD, MS; David K. Blough, PhD; Cara L. McDermott, BA; and Carolina M. Reyes, PhD

ABSTRACT

BACKGROUND: Administrative claims are readily available, but their usefulness for identifying persons with non-small cell lung cancer (NSCLC) is relatively unknown, particularly for younger persons and those enrolled in Medicaid.

OBJECTIVES: To determine the sensitivity of ICD-9-CM codes for identifying persons with NSCLC.

METHODS: This was a retrospective analysis of insurance claims records linked to the Surveillance, Epidemiology, and End Results (SEER) cancer registry for the time period January 1, 2002, through December 31, 2005. Persons included in the sample were identified with NSCLC using SEER morphology and histology codes and were enrolled in a commercial health plan, Medicaid, or Medicare fee-for-service health plans in Washington State. The outcome measure was sensitivity, defined as the percentage of SEER-identified patients who were accurately identified as NSCLC cases using ICD-9-CM diagnoses (162.2, 162.3, 162.4, 162.5, 162.8, 162.9, or 231.2) recorded in any claim field in administrative claims data. We examined the influence of varying the number and timing of administrative codes in relation to the SEER cancer diagnosis date. In multivariate models, we examined the influence of age, sex, and comorbidity on sensitivity.

RESULTS: The sensitivity of 1 medical claim including at least 1 ICD-9-CM code for identifying NSCLC within 60 days of diagnosis as documented in the SEER registry was 51.1% for Medicaid, 87.7% for Medicare, and 99.4% for commercial plan members. Sensitivity can improve at the expense of identifying a portion of patients who are 3 or more months from their true diagnosis date. In multivariate models, age, race, and noncancer comorbidity but not gender significantly influenced sensitivity.

CONCLUSIONS: Administrative claims are sensitive for identifying patients with new NSCLC in the commercial and Medicare plans. For Medicaid patients, linkage with cancer registry records is needed to conduct studies using administrative claims.

J Manag Care Pharm. 2009;15(8):659-68

What is already known about this subject

- Algorithms using administrative claims can vary in accuracy for identifying diseases such as cancer.
- Most prior analyses of claims accuracy have been conducted for a specific cohort in a single health plan.
- There are very few analyses of claims accuracy in Medicaid and private health plans.

What this study adds

- The sensitivity of at least 1 ICD-9-CM code in any field of administrative claims for identifying non-small cell lung cancer (NSCLC) patients within 60 days of diagnosis as documented in the Surveillance, Epidemiology, and End Results (SEER) registry was 51.1% for Medicaid, 88.7% for Medicare, and 99.4% for commercial plan members; the sensitivity for at least 2 ICD-9-CM codes was 39.6% for Medicaid, 86.2% for Medicare fee-for-service, and 97.8% for commercial plan members. Specificity may be important to researchers who wish to avoid cases in which ICD-9-CM codes are falsely positive, but specificity could not be examined in this study due to data agreements with the health plans.
- Among Medicaid enrollees, the sensitivity of the codes was significantly higher for younger persons than for those older than aged 75 years, for nonwhites compared with whites, and significantly lower for those with no comorbidity compared with those with 1 or more comorbidities.
- Among Medicare fee-for-service enrollees with NSCLC, sensitivity was significantly lower for female gender, persons aged 55 years or younger, nonwhites, and persons with no comorbidities.
- Stage of disease might be an important factor to consider when analyzing sensitivity, but this additional analysis was not performed.

The cornerstone of many patterns and cost-of-care studies in cancer are algorithms that use administrative claims data from health insurance plans to identify persons with the cancer of interest.[1-3] Numerous studies have evaluated the accuracy of algorithms for identifying incident cases of breast cancers, particularly among Medicare-eligible women.[3-5] Studies comparing the accuracy of administrative codes for lung cancer compared with cancer registry records among Medicare-eligible patients have found sensitivities of administrative codes ranging from 56% to 90%.[5-7] Knowledge of administrative code sensitivity may facilitate future database and claims research, for example, with research conducted in geographic areas where linkage to clinical data such as medical records or a cancer registry—such as the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER)—is not possible or not feasible.

While these studies focused on Medicare-eligible patients, nearly one-third of lung cancer patients newly diagnosed each

year in the United States are younger than age 65 at the time of diagnosis.[8] The relative accuracy of algorithms using administrative claims to identify incident cases of cancer in Medicaid and private health plans is relatively unknown.

Lung cancer cases may be difficult to identify using administrative claims. Many patients, particularly the elderly, do not receive treatment, making reliance on certain administrative claims codes problematic.[9] In addition, timing of codes relative to the actual point of diagnosis is important for many studies, particularly those seeking to separate diagnostic costs from treatment costs.

With these issues in mind, the purpose of this study was to estimate relative sensitivity of claims for identifying persons with non-small cell lung cancer (NSCLC) in 3 health insurance plans: Medicare, Medicaid, and a private insurer serving persons younger than age 65. We sought to determine the timing of administrative codes in relation to the cancer diagnosis date, as established by cancer registry records. We also sought to examine whether age, race, gender, and other illnesses alter the accuracy of codes across plans.

This research received approval from the Washington State Institutional Review Board (Department of Social and Health Services project application number D-053108-S, "Development of a Claims-Based Algorithm to Identify Incident Cases of Non-Small Cell Lung Cancer").

Methods

Patient-level data obtained from the SEER Puget Sound registry were merged with health care claims from 3 health insurers: Medicare, Washington State Medicaid, and Regence Blue Shield. The SEER records provided patient information regarding tumor characteristics, stage at diagnosis, and survival. Demographic information, such as age, gender, and race, was also obtained from SEER registry records. Health insurance status, comorbidity information, and health system utilization were based on insurance enrollment and administrative claims from the 3 payers.

The SEER-Puget Sound registry, established in 1974 under contract with the federal SEER program, provides high-quality data on the incidence, treatment, and follow-up on newly diagnosed cancers occurring in residents of 13 counties in northwest Washington State.[10] Information on cancer cases is obtained by SEER from hospitals, outpatient surgical centers, pathology laboratories, clinician offices, and death certificates.

Regence Blue Shield is a private nonprofit health insurer providing coverage to more than 1 million Washington State residents.[11] The Medicaid program provides health insurance for approximately 420,000 low-income beneficiaries in Washington State.[12] The Medicare program provides coverage for persons aged 65 and older, persons less than 65 years of age with certain disabilities, and persons of all ages with end-stage renal disease.[13] Our analysis includes only fee-for-service Medicare beneficiaries, as individual claims are not submitted to Medicare risk-sharing plans. Regence, Medicare, and Medicaid claims contain service-level diagnosis and encounter information for all covered services.

To identify subjects with newly diagnosed NSCLC among people living within the 13 counties covered by the SEER-Puget Sound registry, we cross-linked person-level identifiers (full name, gender, date of birth, and in some cases ZIP code) from each plan's enrollment files with histologically confirmed NSCLC cases identified in the SEER-Puget Sound registry. SEER morphology and histology codes are listed in Table 1. Patients aged 25 and older were included in the database because some patients below the age of 25 may have pediatric cancers; however, these cancers under the age of 25 are extremely rare. Inclusion criteria were as follows: (a) aged 25 or older on the date of diagnosis, defined as the first date of histologically confirmed NSCLC appearing in the SEER database; (b) enrollment in the health plan at the SEER date of diagnosis; and (c) NSCLC diagnosis between January 1, 2002, and December 31, 2005. Patients were excluded if they had other malignancies previously recorded at any time in SEER or did not have complete insurance claims records, including incomplete Medicare claims records due to dropping Part B insurance or entering a Medicare HMO at any time during follow-up. Patients' claims were searched for 12 months post-SEER diagnosis date or until date of death, whichever occurred first. This aggregation of claims allowed for standardization of the database.

Using an algorithm developed by Klabunde et al. (2000), a noncancer comorbidity score (based on a count of specific comorbidities) was computed for each patient enrolled in Regence Blue Shield or Medicare based on claims observed in the year prior to SEER diagnosis date.[14] Because patients were commonly enrolled in Medicaid at or shortly after their cancer diagnosis, we constructed comorbidity scores for this population using claims records from the point of enrollment.

Using the SEER cancer registry records as the gold standard, we tested the sensitivity of *International Classification of Diseases, Ninth Revision, Clinical Modification* (ICD-9-CM) codes in any field (Table 1) to identify incident cases of NSCLC. For those with more than 1 ICD-9-CM code, we identified the initial date that 1 of these codes appeared and compared it with the diagnosis date recorded in SEER. The insurer data we had were obtained through a request of claims data for cancer patients identified by SEER. Any false positives in the insurer data would not have appeared in the SEER data; therefore, they would not have been requested from the insurer. For this reason, a specificity measure could not be calculated.

If SEER did not record the diagnosis day (i.e., only month and year), we assigned a diagnosis date of the first day of the diagnosis month, pursuant to our common method for these SEER records. The date of record of administrative codes is known to vary in relation to the service date and the date that a condition appears in clinical records.[15,16] To address the potential impact of this issue on sensitivity, we defined several different time periods to

TABLE 1 SEER and ICD-9-CM Codes Used to Identify Patients with Non-Small Cell Lung Cancer

SEER Morphology Codes	Description
C34.0	Main bronchus (including carina, hilus of lung)
C34.1	Upper lobe (including lingula), lung
C34.2	Middle lung
C34.3	Lower lobe, lung
C34.8	Overlapping lesion of lung
C34.9	Lung, NOS
SEER Histology Codes	**Description**
8000	Malignant neoplasm, NOS
8001	Malignant tumor cells
8010	Carcinoma, NOS
8012	Large cell carcinoma, NOS
8020	Carcinoma, undifferentiated, NOS
8021	Carcinoma, anaplastic, NOS
8022	Pleomorphic carcinoma
8033	Sarcomatoid carcinoma
8041	Small cell carcinoma, NOS
8046	Non-small cell cancer, NOS
8070	Squamous cell carcinoma, NOS
8140	Adenocarcinoma, NOS
8240	Neuroendocrine carcinoma, NOS
8246	Carcinoid, NOS
ICD-9-CM Code[a]	**Description**
162.2	Malignant neoplasm of main bronchus, carina, hilus of lung
162.3	Malignant neoplasm of upper lobe, bronchus, or lung
162.4	Malignant neoplasm of middle lobe, bronchus, or lung
162.5	Malignant neoplasm of lower lobe, bronchus, or lung
162.8	Malignant neoplasm of other parts of bronchus or lung; Malignant neoplasm of contiguous or overlapping sites of bronchus or lung; point of origin undetermined
162.9	Malignant neoplasm of bronchus and lung, unspecified
231.2	Carcinoma in situ, bronchus and lung, carina, hilus of lung

[a]*ICD-9-CM codes are not available for histology.*

ICD-9-CM = International Classification of Diseases, Ninth Revision, Clinical Modification; NOS = not otherwise specified; SEER = Surveillance, Epidemiology, and End Results.

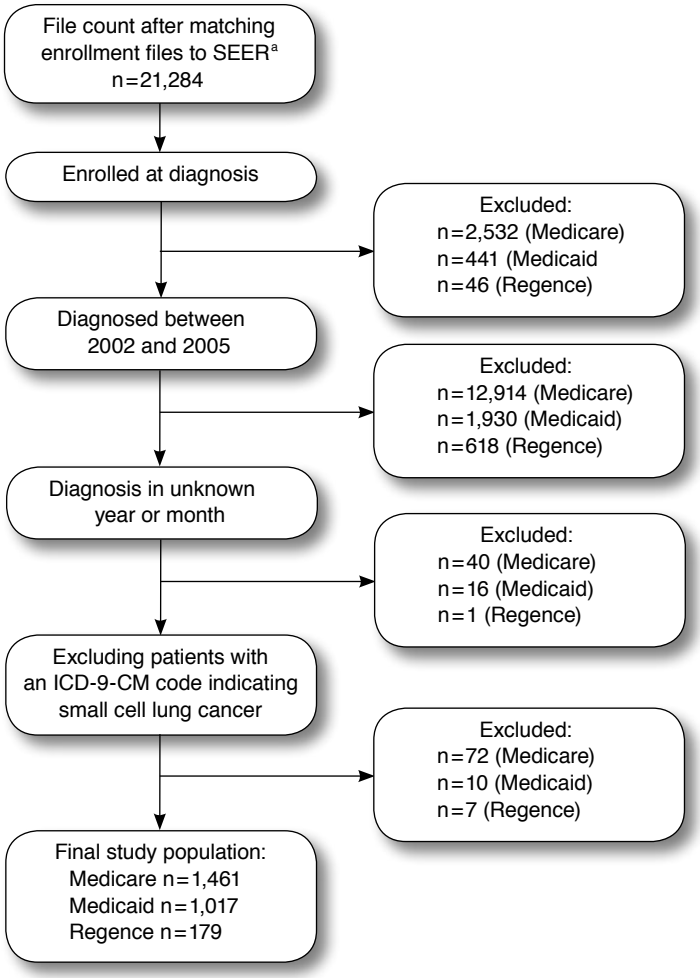
search for ICD-9-CM codes in relation to the SEER-recorded date of diagnosis (in days): –30 to 30, –30 to 60; –30 to 90; 0 to 30; 0 to 60; 0 to 90; 0 to 120. Sensitivity was calculated for each interval. The sensitivity of the claims codes increases with the length of time between the service date and the diagnosis date. Therefore, for newly diagnosed cases, the claims data may not be sensitive enough to be useful. The analysis presented exhaustively examines the effect of different lag times. When multiple claims for 1 patient were recorded, the date of the first claim was used.

We calculated sensitivity using 1 ICD-9-CM code versus 2 separately recorded ICD-9-CM codes within each time period. Each ICD-9-CM code recorded had a service date. Some patients had more than 1 lung cancer code recorded; it made no difference whether or not they had the same service date. We computed sensitivity for patients across all health plans and stratified by individual health plan. Some patients were enrolled in 2 health plans in our study (e.g., Medicare and Regence Blue Shield). These patients were assigned to the plan that had the greatest volume of cancer claims over the period of interest. For our analysis, all administrative claims from both plans were added to that individual's record.

We created multivariate analyses of factors that could influence sensitivity, using weighted least squares, treating the sensitivity within each covariate class as the outcome. Weights are the number of observations in the covariate classes. Weighting is necessary in the linear model, since we are directly modeling sensitivity, a proportion. The variance of each proportion depends on the number of observations that go into that proportion as well as the value of the proportion itself. The method we used was originated by Grizzle, Starmer, and Koch (1969)[17] and is often referred to as the GSK method. We used the CATMOD procedure in SAS, v9.2 (SAS Institute Inc., Cary, NC) to implement the method. Results are significant if $P < 0.05$.

Covariates included age (in years, categorized as 55 or

FIGURE 1 Application of Exclusion and Inclusion Criteria to Create Database for Analysis



[a]*Excluding patients with other malignancies recorded at any time in SEER or lacking complete claims records. See SEER codes in Table 1.*

ICD-9-CM = International Classification of Diseases, Ninth Revision, Clinical Modification; SEER = Surveillance, Epidemiology, and End Results.

younger, 56 to 75, and greater than 75), gender, race (white or nonwhite [race is available in SEER data]), and comorbidities as defined by the Klabunde method (0, 1, or more than 1). These are included in the regression model as main effects. A priori, we had no hypotheses of interactions among the predictor variables. However, by including all interactions in the model, we obtained a fit of the so-called saturated model. This model has as many parameters as covariate classes. Thus, it fits the data perfectly in the sense that the predicted values from the saturated model are identical to the observed covariate class sensitivities. This approach is similar to fitting a line to 2 data points or a parabola to 3 data points. The difference in fit (via a Wald test) between the saturated model and the main effects only model provided an assessment of the lack of fit of the main effects model. Lack of a statistically significant difference between the saturated model and the main effects model means that the latter model fits well.

TABLE 2 Characteristics of Study Subjects with Non-Small Cell Lung Cancer, as Recorded by SEER

Characteristic	Regence	Medicaid	Medicare	All Plans
Number	179	1,017	1,461	2,657
Mean age at diagnosis[a]	61.8	65.7	74.6	70.3
Standard deviation	10.3	12.2	7.6	10.9
Male gender (%)[a]	57.0%	49.7%	62.5%	57.2%
Race[a]				
White (%)	93.3%	78.7%	93.2%	87.6%
Nonwhite (%)	6.7%	21.3%	6.8%	12.4%
Comorbidity score[a] mean [SD]	0.7 [1.2]	1.0 [1.6]	1.8 [1.8]	1.5 [1.7]

[a]*P < 0.001 for these differences among the health plans. Multivariate analysis, using weighted least squares estimate was performed; Medicare was the reference plan. The comorbidity measure is the Klabunde comorbidity algorithm.*[14]

SEER = Surveillance, Epidemiology, and End Results.

Results

After linking SEER records with health plan claims and applying exclusion criteria (Figure 1), a total of 2,657 persons enrolled in the 3 health plans were diagnosed with NSCLC between 2002 and 2005. The average age was 70.3 (standard deviation [SD] = 10.9); 42.8% were female (Table 2). The greatest proportion of nonwhite cancer patients were enrolled in Medicaid. Medicare patients had the highest average comorbidity score at the time of diagnosis; Regence Blue Shield patients had the lowest.

The overall sensitivity of ICD-9-CM codes varied substantially by plan type (Table 3). Algorithm sensitivity was lowest for Medicaid enrollees and highest for Regence enrollees. Sensitivity was lower when 2 separate ICD-9-CM codes were required to indicate a cancer diagnosis. Stratified by time period in relation to diagnosis, sensitivity generally increased over wider time horizons, suggesting that some NSCLC patients are found by administrative coding months after the diagnosis date appearing in SEER.

Using the diagnosis date as recorded by SEER compared with a 0- to 30-day time horizon, the percentage of additional cases detected by ICD-9-CM codes over the additional time horizon at 90 days, for example, was 12% to 17% in Regence Blue Shield, 14% to 19% in Medicaid, and 7% to 9% in Medicare, depending on whether 1 or 2 separate ICD-9-CM codes are used to identify an individual as having NSCLC. The highest sensitivities included administrative codes up to 120 days following the SEER diagnosis date. Including ICD-9-CM codes that appeared 30 days prior to the SEER diagnosis date had little impact on sensitivity compared with only including codes that appeared

TABLE 3 *ICD-9-CM Algorithms for Identifying Incident NSCLC Cases*

Claims Observation Period Relative to SEER Date of Diagnosis							
Regence Blue Shield	**–30 to 30**	**–30 to 60**	**–30 to 90**	**0 to 30**	**0 to 60**	**0 to 90**	**0 to 120**
Number with at least 1 ICD-9-CM code	156	178	178	156	178	178	179
Sensitivity (%)	87.2	99.4	99.4	87.2	99.4	99.4	100.0
Number with at least 2 ICD-9-CM codes	146	175	177	146	175	177	177
Sensitivity (%)	81.6	97.8	98.9	81.6	97.8	98.9	98.9
Medicaid	**–30 to 30**	**–30 to 60**	**–30 to 90**	**0 to 30**	**0 to 60**	**0 to 90**	**0 to 120**
Number with at least 1 ICD-9-CM code	438	525	572	430	520	567	595
Sensitivity (%)	43.1	51.6	56.2	42.3	51.1	55.8	58.5
Number with at least 2 ICD-9-CM codes	289	409	483	282	403	478	507
Sensitivity (%)	28.4	40.2	47.5	27.7	39.6	47.0	49.9
Medicare	**–30 to 30**	**–30 to 60**	**–30 to 90**	**0 to 30**	**0 to 60**	**0 to 90**	**0 to 120**
Number with at least 1 ICD-9-CM code	1,218	1,296	1,321	1,204	1,282	1,307	1,312
Sensitivity (%)	83.4	88.7	90.4	82.4	87.7	89.5	89.8
Number with at least 2 ICD-9-CM codes	1,162	1,269	1,294	1,149	1,260	1,285	1,293
Sensitivity (%)	79.5	86.9	88.6	78.6	86.2	88.0	88.5

ICD-9-CM = International Classification of Diseases, Ninth Revision, Clinical Modification; NSCLC = non-small cell lung cancer; SEER = Surveillance, Epidemiology, and End Results.

on or after the SEER-recorded diagnosis date.

The weighted least squares multivariate regression models showed good fit overall for the Medicare and Medicaid patient groups ($P = 0.60$ and 0.08, respectively), but because of the small number of cases observed in the Regence patient group, the model failed to produce meaningful estimates at all. Considering the 0- to 60- , 0- to 90- , and 0- to 120-day time periods, among those enrolled in Medicaid the sensitivity of the codes was significantly higher for younger persons than for those older than aged 75 years and for nonwhites compared with whites. Sensitivity was significantly lower for those with no comorbidity compared with those with 1 or more comorbidities. With respect to the association between sensitivity and gender, we were not able to reject the null hypothesis.

Among Medicare enrollees with NSCLC, sensitivity was significantly lower for female gender, persons aged 55 years or younger, nonwhites, and persons with no comorbidities. We created regression models for the 0- to 30- , 0- to 60- , 0- to 90- , and 0- to 120-day time periods. There were fewer significant associations for the 30-day time period, but little difference between the 60- , 90- , and 120-day time periods.

Figure 2 shows the adjusted sensitivity values for Medicare and Medicaid enrollees considering the different time horizons. The overall pattern of coefficient estimates with each plan is quite similar for the various time windows. We show the estimated coefficients and their standard errors in Table 4 for the 120-day time period. For the Medicare enrollees, there is a significant difference in gender: women show a 5% decrease in sensitivity relative to men. Those enrollees aged 55 years and younger show a 25% decrease in sensitivity relative to those over 75, and nonwhites show a 9% decrease relative to whites. Those with no comorbidities show a 5% reduction in sensitivity relative to those with 2 or more.

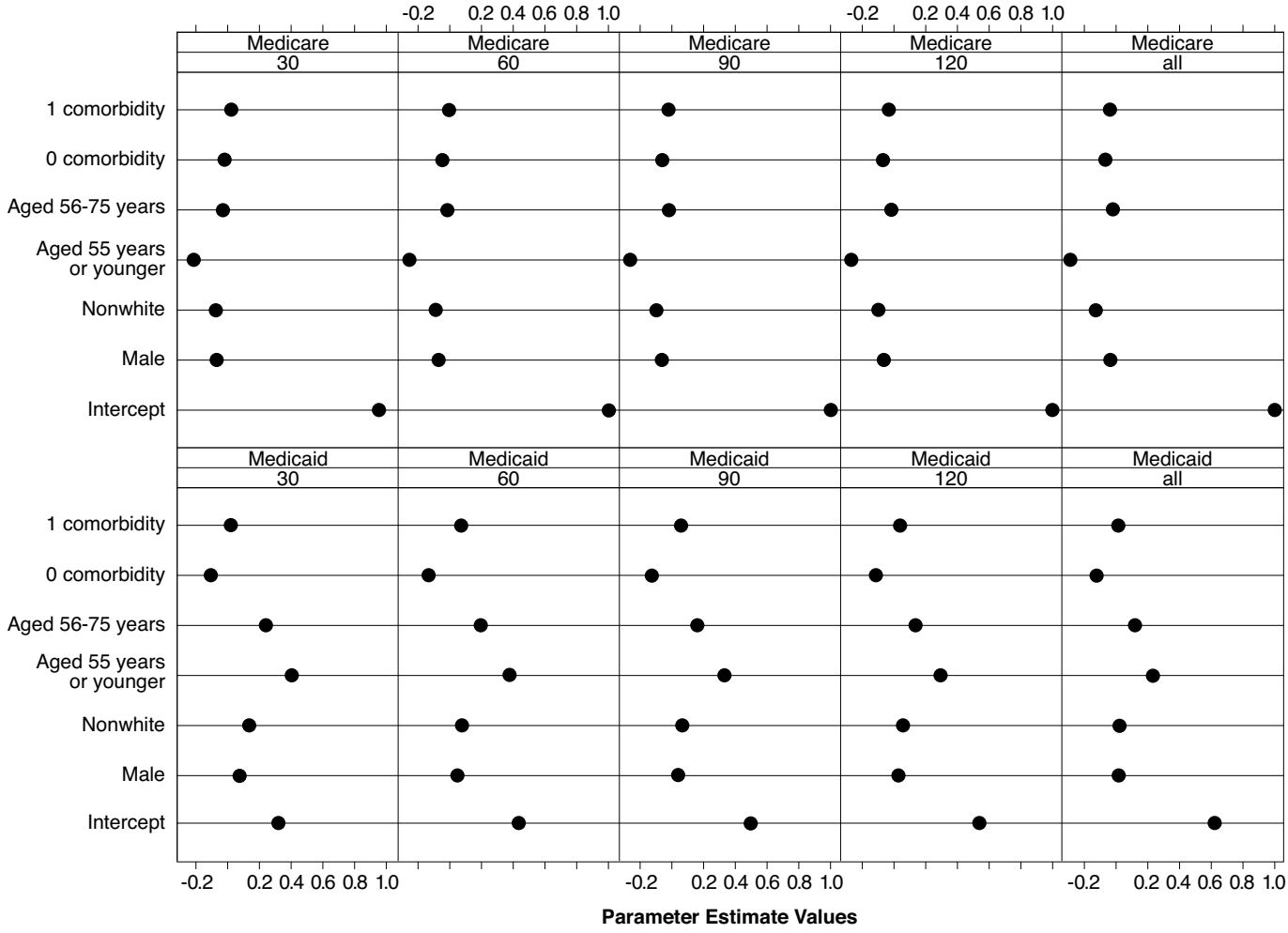
For Medicaid enrollees, those 55 years of age or less show a 31% increase in sensitivity relative to those older than 75, while those aged 56 years to 75 years show a 15% increase. Those with no comorbidities show a decrease in sensitivity of 10% relative to those with 2 or more comorbidities. These regression model results, along with standard errors and P values, are shown in Table 4.

Discussion

Conducting cancer outcomes research using administrative claims records requires accurate identification of persons with the cancer of interest. In this evaluation of persons with histologically confirmed NSCLC in 3 health insurance plans in Washington State, we found high overall sensitivity when using a single ICD-9-CM code to identify persons with NSCLC while enrolled in Medicare and a commercial insurance plan, but modest sensitivity among persons enrolled in Medicaid. If our results are applied to other commercial and regional Medicare plans, health services researchers may be able to use a relatively simple algorithm of a single ICD-9-CM code to identify most persons with NSCLC, although use of a single ICD-9-CM code may contribute to false positives in the absence of linkage to a SEER registry. Use of a single code may save resources with less programming time while increasing potential sample size of future studies.

If timing in relation to the true diagnosis date is critical for a particular analysis (e.g., to determine relationship of date of diagnosis to date of initial treatment), these analyses suggest that health plan type may be an important factor. Over 83% of Medicare NSCLC cases and 87% of commercial plan NSCLC

FIGURE 2 Plot of Parameter Values[a] from the Weighted Least Squares Multivariate Analyses Showing Sensitivity of a Single ICD-9-CM Code for Identifying Persons with NSCLC by Plan Type (Medicare, Medicaid) and Time from Diagnosis (30, 60, 90, 120 Days and a Match for Any of These Time Windows, Labeled as "All")



[a]*For gender, the reference group is female; for age, the reference group is aged older than 75 years; for race, the reference group is white; for comorbidities, the reference group is 2 or more.*

ICD-9-CM = International Classification of Diseases, Ninth Revision, Clinical Modification; NSCLC = non-small cell lung cancer.

cases were identified within 30 days of the SEER diagnosis date, but fewer were identified in Medicaid. Some lung cancer patients are not treated for their cancer, as the result of being too ill to withstand treatment or choosing not to be treated. Some may also die after a single treatment or discontinue treatment. Therefore the ≥2 code cohort will be less numerous than the 1 code cohort.

The sensitivity of ICD-9-CM codes for identifying NSCLC cases was substantially inferior for Medicaid compared with the other 2 health plans. Medicaid provides coverage to a heterogeneous group of patients, many of whom enroll only after being newly diagnosed with cancer. Furthermore, gaps in enrollment and disenrollment shortly after enrolling in Medicaid appear to be common.[18] We postulate that these breaks are the primary reason why ICD-9-CM codes have limited sensitivity for Medicaid enrollees with NSCLC. Other issues unique to Medicaid populations versus privately enrolled or Medicare-enrolled patients might include lack of timely follow-up after an initial evaluation due to access barriers or perhaps differences in how providers code visits for Medicaid patients versus those with other types of insurance.

TABLE 4 Estimated Regression Coefficients from the Weighted Least Squares Regression Model, 120-Day Window

Medicare								
Source	Analysis of Variance			Analysis of Weighted Least Squares Estimates				
	DF	Chi-Square	Pr > Chi-Square		Estimate	Standard Error	Chi-Square	Pr > Chi-Square
Intercept	1	8768.72	<0.0001		0.9921	0.0106	8768.72	<0.0001
Gender[a]	1	17.76	<0.0001	1	-0.0484	0.0115	17.76	<0.0001
Age category	2	9.60	0.0082	56-75	-0.0040	0.0113	0.12	0.7241
				≤55	-0.2516	0.0813	9.58	0.0020
Race[b]	1	6.55	0.0105	0	-0.0860	0.0336	6.55	0.0105
Comorbidity	2	11.98	0.0025	0	-0.0532	0.0154	11.91	0.0006
				1	-0.0155	0.0132	1.37	0.2412
Medicaid								
Source	Analysis of Variance			Analysis of Weighted Least Squares Estimates				
	DF	Chi-Square	Pr > Chi-Square		Estimate	Standard Error	Chi-Square	Pr > Chi-Square
Intercept	1	137.42	<0.0001		0.5418	0.0462	137.42	<0.0001
Gender[a]	1	1.86	0.1728	1	0.0423	0.0310	1.86	0.1728
Age category	2	48.83	<0.0001	56-75	0.1475	0.0424	12.09	0.0005
				≤55	0.3050	0.0452	45.58	<0.0001
Race[b]	1	3.69	0.0548	0	0.0708	0.0369	3.69	0.0548
Comorbidity	2	15.16	0.0005	0	-0.0984	0.0409	5.80	0.0160
				1	0.0517	0.0346	2.23	0.1350

[a]0 = male, 1 = female
[b]0 = white, 1 = nonwhite
DF = degrees of freedom.

Among Medicare enrollees, sensitivity was significantly lower for women, younger persons, nonwhites, and those with no comorbidities. It is possible that lung cancer is less suspected in these individuals, thus, less frequently coded. Another possibility is that persons are identified clinically (i.e., in charts) but not recorded in claims because treatments are not initiated. Most lung cancers are diagnosed at advanced stage, and only a minority of patients with advanced stage lung cancer receives treatment for the disease.[8] Those with fewer comorbidities may not be diagnosed because they are less likely to see a physician in general and, thus, have fewer opportunities for a code to be recorded. Among Medicaid enrollees, sensitivity was quite low in general, making interpretation of individual coefficients less useful for decision makers.

Limitations

We note limitations of this study. First, agreements with the respective health plans permitted us to obtain only SEER-confirmed cases that were enrolled in each plan. Thus, we were unable to generate specificity values. Specificity may be important to researchers who wish to avoid cases where ICD-9-CM codes are falsely positive. Second, stage of disease might be an important factor to consider when analyzing sensitivity; however, we did not perform this analysis. Third, the results are restricted to Washington State so may not apply directly to other health plans in other states because of variation in eligibility requirements and regional coding practices.

Conclusion

The sensitivity of administrative claims appears to be high for identifying newly diagnosed NSCLC patients in Medicare and commercial insurance in as little as 60 days following the clinical diagnosis date as recorded by SEER. Identifying Medicaid enrollees is problematic most likely because of cancer-specific enrollment and high disenrollment rates shortly after cancer diagnosis. Age at diagnosis, race, and comorbidity but not gender may significantly influence sensitivity.

Authors

SCOTT D. RAMSEY, MD, PhD, is Full Member; JOHN F. SCOGGINS, PhD, MS, is Postdoctoral Fellow; and CARA L. MCDERMOTT, BA, is Project Coordinator, Fred Hutchinson Cancer Research Center, Seattle, Washington. DAVID K. BLOUGH, PhD, is Research Associate Professor, School of Pharmacy, University of Washington, Seattle, Washington. CAROLINA M. REYES, PhD, is Senior Health Economist, Genentech, Inc., San Francisco, California.

AUTHOR CORRESPONDENCE: Scott D. Ramsey, MD, PhD, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N., M3-B232, Seattle, WA 98109. Tel.:206.667.7846; Fax: 206.667.5977; Email: sramsey@fhcrc.org.

DISCLOSURES

This research was funded by Genentech, Inc., and Carolina Reyes is an employee of Genentech. The authors report that there is no relationship that could be construed as an actual, potential, or apparent conflict of interest with regard to the subject of this manuscript. An abstract based on the research described in this article was accepted and presented at the Academy of Managed Care Pharmacy meeting in Orlando, Florida, April 15-18, 2009.

Ramsey was primarily responsible for the study concept and design, with assistance from Blough and Reyes. The data were collected by Ramsey and McDermott and interpreted by Ramsey, Scoggins, Blough, and Reyes. The manuscript was written primarily by Ramsey with assistance from Scoggins and McDermott. Ramsey made the majority of the revisions with assistance from Blough and McDermott.

REFERENCES

1. Rolnick SJ, Hart G, Barton MB, et al. Comparing breast cancer case identification using HMO computerized diagnostic data and SEER data. *Am J Manag Care*. 2004;10(4):257-62. Available at: http://www.ajmc.com/media/pdf/AJMC2004AprRolnick257_262.pdf. Accessed September 20, 2009.

2. Gold HT, Do HT. Evaluation of three algorithms to identify incident breast cancer in Medicare claims data. *Health Serv Res*. 2007;42(5):2056-69. Abstract available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17850533. Accessed September 20, 2009.

3. Cooper GS, Yuan Z, Stange KC, Dennis LK, Amini SB, Rimm AA. The sensitivity of Medicare claims data for case ascertainment of six common cancers. *Med Care*. 1999;37(5):436-44. Abstract available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10335746. Accessed June 27, 2009.

4. Koroukian SM, Cooper GS, Rimm AA. Ability of Medicaid claims data to identify incident cases of breast cancer in the Ohio Medicaid population. *Health Serv Res*. 2003;38(3):947-60. Available at: http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=1360924&blobtype=pdf. Accessed September 20, 2009.

5. McClish DK, Penberthy L, Whittemore M, et al. Ability of Medicare claims data and cancer registries to identify cancer cases and treatment. *Am J Epidemiol*. 1997;145(3):227-33. Available at: http://aje.oxfordjournals.org/cgi/reprint/145/3/227. Accessed September 20, 2009.

6. Setoguchi S, Solomon DH, Glynn RJ, Cook EF, Levin R, Schneeweiss S. Agreement of diagnosis and its date for hematologic malignancies and solid tumors between Medicare claims and cancer registry data. *Cancer Causes Control*. 2007;18(5):561-69. Abstract available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17447148. Accessed September 20, 2009.

7. McBean AM, Warren JL, Babish JD. Measuring the incidence of cancer in elderly Americans using Medicare claims data. *Cancer*. 1994;73(9):2417-25. Abstract available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=8168045. Accessed September 20, 2009.

8. National Cancer Institute. SEER Stat Fact Sheets. Available at: http://www.seer.cancer.gov/statfacts/html/lungb.html. Accessed September 20, 2009.

9. Whittle J, Steinberg EP, Anderson GF, Herbert R. Accuracy of Medicare claims data for estimation of cancer incidence and resection rates among elderly Americans. *Med Care*. 1991;29(12):1226-36. Abstract available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=1745080. Accessed September 20, 2009.

10. Fred Hutchinson Cancer Research Center. Cancer Surveillance System. 2008. CSS Reports, Publications, and FAQs. Available at: http://www.fhcrc.org/science/phs/css/publications.html. Accessed September 20, 2009.

11. Regence Blue Shield. Annual reports. Available at: http://www.regence.com/about/annualReport/annual-report.jsp. Accessed September 20, 2009.

12. Washington State Health Recovery Services Administration. 2008. People Enrolled in DSHS Medical Programs by County, February 2008. Available at: http://hrsa.dshs.wa.gov/News/EnrollmentFigures/PeopleEnrolledinDSHSMedicalProgramsbyCounty.xls. Accessed September 20, 2009.

13. Centers for Medicare and Medicaid Services. 2008. Medicare Program--General Information. Available at: http://www.cms.hhs.gov/MedicareGenInfo/. Accessed September 20, 2009.

14. Klabunde CN, Potosky AL, Legler JM, Warren JL. Development of a comorbidity index using physician claims data. *J Clin Epidemiol*. 2000;53(12):1258-67. Abstract available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11146273. Accessed September 20, 2009.

15. Brackley ME, Penning MJ, Lesperance ML. In the absence of cancer registry data, is it sensible to assess incidence using hospital separation records? *Int J Equity Health*. 2006;5:12. Available at: http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=1613240&blobtype=pdf. Accessed September 20, 2009.

16. Kern EF, Maney M, Miller DR, et al. Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease in diabetes. *Health Serv Res*. 2006;41(2):564-80. Available at: http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=1702507&blobtype=pdf. Accessed September 20, 2009.

17. Grizzle JE, Starmer CF, Koch GG. Analysis of categorical data by linear models. *Biometrics*. 1969;25(3):489-504.

18. Ramsey SD, Zeliadt SB, Richardson LC, et al. Disenrollment from Medicaid after recent cancer diagnosis. *Med Care*. 2008;46(1):49-57. Abstract available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18162855. Accessed September 20, 2009.

APPENDIX 1 *Administrative Codes Used to Identify NSCLC, First-Line Chemotherapy, G-CSF, and Infection Therapy*

Measure	ICD-O-2 Histology[a]	HCPCS/CPT
Diagnosis		
Large cell carcinoma	8012/3	
Squamous cell carcinoma	8070/3	
Squamous cell carcinoma, keratinizing	8071/3	
Squamous cell carcinoma, large cell, nonkeratinizing	8072/3	
Adenocarcinoma	8140/3	
Bronchiolo-alveolar adenocarcinoma	8250/3	
Mucinous adenocarcinoma	8480/3	
Mucin-producing adenocarcinoma	8481/3	
Signet ring cell carcinoma	8490/3	
Adenoquamous carcinoma	8560/3	
Adenocarcinoma with squamous metaplasia	8570/3	
First-line chemotherapy		
Cisplatin		C9418, J9060, J9062
Carboplatin		J9045
Paclitaxel		C9127, C9431, J9264, J9265
Docetaxel		J9170
Gemcitabine		J9201
Vinorelbine		C9440, J9390
Irinotecan		J9206
Etoposide		C9414, C9425, J8560, J9181, J9182
Vinblastine		J9360
Bevacizumab		C9214, J9035, S0116
Pemetrexed		C9213, J9305
G-CSF		
Filgrastim		J1440, J1441
Pegfilgrastim		C9119, J2505, Q4053, S0135
Diagnostic Testing		
Complete blood count		85025, 85027
Urine culture		87086, 87087, 87088
Chest x-ray		71010, 71015, 71020, 71021, 71023, 71030, 71034, 71035
Blood culture		87040
Throat culture		87060, 87081
Stool culture		87045, 87046
Infection therapy		
Intravenous infusion for therapy/diagnosis		90780, 90781
Intramuscular injection of antibiotic		90788
Home infusion therapy, antibiotic, antiviral, or antifungal therapy		S9494, S9497, S9500, S9501, S9502, S9503, S9504

[a]*NSCLC was identified using ICD-O-2 histology codes used in the SEER database.*

CPT = Current Procedural Terminology; G-CSF = granulocyte-colony stimulating factor; HCPCS = Healthcare Common Procedure Coding System; NSCLC = non-small cell lung cancer; SEER = Surveillance, Epidemiology and End Results.

APPENDIX 2 SEER Variables for Identification of Stage IIIB NSCLC[a]

Variable Description	E10EX1	E10DN1
Tumor extension		
Carina; trachea; esophagus Mediastinum, extrapulmonary or NOS Major blood vessel(s): Pulmonary artery or vein; superior vena cava (SVC syndrome); aorta; azygos vein Nerve(s): Recurrent laryngeal (vocal cord paralysis); vagus; phrenic; cervical sympathetic (Horner's syndrome)	70	
Heart, visceral pericardium	71	
Malignant pleural effusion Pleural effusion, NOS	72	
Sternum Vertebra(e) Skeletal muscle Skin of chest	75	
Pericardial effusion, NOS; malignant pericardial effusion	79	
Regional lymph nodes Contralateral hilar or mediastinal (including bilateral) Supraclavicular (transverse cervical), ipsilateral or contralateral Scalene, ipsilateral or contralateral		6
Distant lymph nodes		7

[a]*Patients identified as Stage IIIB if 1 of these codes for tumor extension of lymph node involvement was present in SEER.*

NSCLC = non-small cell lung cancer; NOS = not otherwise specified; SEER = Surveillance, Epidemiology and End Results.