Cancer Causes Control (2007) 18:561–569
DOI 10.1007/s10552-007-0131-1

ORIGINAL PAPER

Agreement of diagnosis and its date for hematologic malignancies and solid tumors between medicare claims and cancer registry data

Soko Setoguchi · Daniel H. Solomon · Robert J. Glynn · E. Francis Cook · Raisa Levin · Sebastian Schneeweiss

Received: 17 November 2006 / Accepted: 16 February 2007 / Published online: 19 April 2007

Abstract

Purpose Claims data may be a suitable source studying associations between drugs and cancer. However, linkage between cancer registry and claims data including pharmacy-dispensing information is not always available. We examined the accuracy of claims-based definitions of incident cancers and their date of diagnosis.

Methods Four claims-based definitions were developed to identify incident leukemia, lymphoma, lung, colorectal, stomach, and breast cancer. We identified a cohort of subjects aged ≥65 (1997–2000) from Pennsylvania Medicare and drug benefit program data linked with the state cancer registry. We calculated sensitivity, specificity, and positive predictive values of the claims-based definitions using registry as the gold standard. We further assessed the agreement between diagnosis dates from two data sources.

Results All definitions had very high specificity (≥98%), while sensitivity varied between 40% and 90%. Test characteristics did not vary systematically by age groups. The date of first diagnosis according to Medicare data tended to be later than the date recorded in the registry data except for breast cancer. The differences in dates of first diagnosis were within 14 days for 75% to 88% of the cases. Bias due to outcome misclassification of our claims-based definition of cancer was minimal in our example of a cohort study.

Conclusions Claims data can identify incident hematologic malignancies and solid tumors with very high specificity with sufficient agreement in the date of first diagnosis. The impact of bias due to outcome misclassification and thus the usefulness of claims-based cancer definitions as cancer outcome markers in etiologic studies need to be assessed for each study setting.

Keywords Incident cancer · Medicare claims · Cancer registry · Date of diagnosis · Agreement · Senstivity · Sepcificity · Positive predictive value · Misclassification

S. Setoguchi (✉) · D. H. Solomon · R. J. Glynn · R. Levin · S. Schneeweiss
Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA
e-mail: ssetoguchi@partners.org

D. H. Solomon
Division of Rheumatology, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

E. F. Cook
Division of General Internal Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

E. F. Cook · S. Schneeweiss
Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA

Background

Health care utilization databases such as claims files from Medicaid, Medicare linked with pharmacy prescription programs, or large health maintenance organizations (HMOs) have been used to study the safety of drugs in pharmacoepidemiologic studies. New biologic immunomodifying drugs such as tumor necrosis factor (TNF)-α antagonists have raised concerns about an increased risk of cancers, especially lymphoproliferative malignancies [1, 2]. Large population-based datasets are required to examine the effect of infrequent exposures (e.g., TNF-α antagonists) on rare outcomes (e.g., lymphoma). Although the

Surveillance Epidemiology and End Results (SEER)-Medicare dataset can provide large numbers of patients with valid cancer diagnosis, the SEER-Medicare dataset does not include prescription drug information. Furthermore, cancer registry data linkable to specific health care utilization data including pharmacy information are not often available, which leaves only health care utilization databases to identify incident cancers.

Previous studies have evaluated the accuracy of cancer diagnoses in Medicare claims data including breast, colorectal, endometrial, lung, pancreatic, and prostate cancers [3–9], but they did not include less frequent cancers such as hematologic malignancies, and the agreement of diagnosis dates between claims data and cancer registry data has not been understood since only month and year of diagnosis are available in SEER data. We sought to develop various claims-based definitions for incident lymphoma and leukemia as well as breast, lung, colorectal, and stomach cancer and assessed the accuracy of these definitions in comparison with registry data including the accuracy of the date of the clinical cancer diagnosis.

Methods

Data sources and study participants

Three data sources were used for this study. Health care utilization data were derived from Medicare claims data linked to pharmacy dispensing data from the Pharmaceutical Assistance Contract for the Elderly (PACE) in Pennsylvania between 1 January 1997 and 31 December 2000. Our gold standard cancer information was Pennsylvania State (PA) Cancer Registry data from 1 January 1989 to 31 December 2000. PACE provides comprehensive pharmacy coverage with only a co-payment of US $6 per prescription. The PA cancer registry is a population-based cancer registry that routinely collects data on patient demographics, date of diagnosis, primary tumor site, morphology, stage at diagnosis, first course of treatment, follow-up for vital status, and survival rates within each stage. The PA cancer registry is certified as ‘‘Gold’’, the highest quality by the North American Association of Central Cancer Registries [10]. The Institutional Review Board of the Brigham and Women's Hospital approved this study, and data use agreements were in place.

We identified a cohort of subjects' age 65 or older who were continuously enrolled in PA Medicare and PACE between 1 January 1997 and 31 December 2000. To ensure subjects' enrollment, we required all subjects to have at least one claim for any service and prescription during each 6-month period until subjects die, or until the study period ends. We also required all subjects to be enrolled, and have no cancer-related claims during the 6 months before 1 January 1997 to exclude subjects currently undergoing treatment for cancer.

Incident case definitions in medicare/PACE

A panel consisting of epidemiologists, health services researchers, and clinical oncologists reviewed and developed four claims data-based definitions of incident cancer using (1) ICD-9 diagnosis codes, (2) Current Procedural Terminology (CPT) codes for screening procedures, surgical procedures, radiation therapy, chemotherapy, and nuclear medicine procedures, and/or (3) National Drug Code (NDC) prescription codes for medications used for cancer treatment available in PACE. The four definitions were developed based on our expectation that we will find some variation in sensitivity and specificity among these definitions. Using four claims data-based definitions (Fig. 1), we identified incident cases of lymphoma and leukemia, as well

Definition 1: Any of the following

>=1 cancer diagnosis + any diagnosis or procedure codes related to complications of cancer or palliative care in two weeks followed by another diagnosis of cancer within 12 months.

≥ 1 diagnostic procedure with biopsy followed by >=2 cancer diagnoses at two different occasions within 12 months (recorded on different dates from the procedures).

≥ 1 cancer diagnosis + any surgery related to cancer during the same hospitalization and/or visit.

≥ 1 cancer diagnosis + any cancer chemotherapy during the same hospitalization and/or visit

≥ 1 cancer diagnosis + any radiation therapy during the same hospitalization and/or visit

≥ 1 cancer diagnosis + hematopoietic cell transplantation during the same hospitalization and/or visit (for leukemia only)

≥ 1 cancer diagnosis + oral chemotherapy dispensing within 2 weeks after the diagnosis

Definition 2:

≥ 2 diagnoses of cancer within 2 months

Definition 3:

Cases defined by using Definition 1 or 2

Definition 4:

≥ 1 diagnosis of cancer

Fig. 1 Health care utilization-based definitions for incident cancers

as breast, colorectal, stomach, and lung cancer diagnosed between 1997 and 2000 in the cohort. Definition 1 is based on combination of diagnoses and procedure codes, Definition 2 and 4 are based on only diagnoses codes, and Definition 3 is defined either as Definition 1 or 2. (Fig. 1) These definitions reflect how researchers typically defined diseases in administrative data. The index date for each case was defined as the earliest date of cancer diagnosis appearing in the health care utilization data. Diagnosis and procedure codes used are available upon request.

Incident case definition in the registry

Incident cancer cases in the registry were those recorded as having lymphoma and leukemia as well as cancers of breast, colorectal, stomach, and lung with the diagnosis date during the study period (1 January 1997–31 December 2000). We considered the cases identified in the registry as the gold standard for validating claims data-based definitions of cancers.

Characteristics of patients

Information on age, gender, race, adjusted net income, and death was obtained from PACE eligibility files and PA cancer registry data files. In addition, information on the stage of the cancer at diagnosis and procedures of diagnostic confirmation were obtained from the registry data.

Data linkage

Participants in the cohort from Medicare/PACE data were linked with the PA cancer registry data using social security number, gender, and date of birth. All person-specific identifiers were removed after successfully linking all the three data sources. Anonymously coded study numbers were used to identify subjects to protect the privacy of program participants.

Statistical analysis

Specificity, sensitivity, and positive predictive value

For the four claims data-based definitions of six types of cancers, we calculated the sensitivity (the number of cancer cases identified by a claims data-based definition that are also identified in the registry divided by the number of all cases with the cancer identified in the registry), specificity (the number of subjects without cancer using claims data-based definition, and the registry divided by the number of all subjects without cancer according to the registry), and PPV (the number of all cases identified by a claims data-based definition that are also identified in the registry divided by the number of cancer cases identified by claims data-based definitions) [11] We also assessed whether these measures varied among age categories (65–74, 75–84, 85+).

Misclassification of prevalent cases as incident cases in health care utilizations data

Contrary to registry information, health care utilization data do not record the onset of diseases. Therefore, we assessed the extent to which prevalent or recurrent cancer cases could be misclassified as incident cases in the health care utilization data depending on a required cancer-free period before the study period. For example, when we required a 6 months cancer-free period, patients had to have no claims with cancer diagnoses during the 6 months prior to 1/1/1997 but had to have at least one claims for any type of health service excluding cancer to ensure health system use. All analyses described above used the default of a 6-month, cancer-free period, but we subsequently varied the period from 0 to 36 months. Sensitivity, specificity and PPV as a function of the duration for the required cancer-free period were calculated with our preferred definition, Definition 3. We chose the preferred definition based on its specificity and relatively high sensitivity. Very high specificity in defining cancers is essential to obtain unbiased ratio estimates for epidemiologic studies assessing the risk of cancer [12]. Relatively high sensitivity is preferred to identify a large proportion of true cases to improve statistical efficiency of estimates, especially when studying rare outcomes.

Date of the onset of cancer

The first date of a cancer diagnosis appeared in the claims data was defined as the incident cancer diagnosis in the claims data. Among subjects who were identified as incident cancer cases by both the registry and Definition 3 in the Medicare claims data, we calculated the difference in days between the cancer diagnosis dates recorded in the two data sources (‘registry diagnosis date’–‘claims diagnosis date’).

Results

Characteristics of study population and cases identified by cancer registry

We identified 157,310 subjects in the cohort who were continuously enrolled in Medicare and PACE and had no cancer-related claims during the 6 months before the study

Table 1 Characteristics of patients enrolled in medicare/PACE who were linked to the PA Cancer Registry between 1997 and 2000

	Total, 6 types (*n* = 6,996[a])	Lung (*n* = 1,810)	Colorectal (*n* = 2,128)	Stomach (*n* = 236)	Female Breast (*n* = 2,004)	Lymphoma (*n* = 629)	Leukemia (*n* = 182)
Gender, n (%)							
Male	1,356 (19.4)	626 (34.6)	461 (21.7)	71 (30.1)	0 (0.0)	136 (21.6)	55 (30.2)
Female	5,640 (80.6)	1,184 (65.4)	1,667 (78.3)	165 (69.9)	2,004 (100.0)	493 (78.4)	127 (69.8)
Age at Diagnosis (mean, SD)	80.0 (6.7)	78.4 (6.3)	81.2 (6.8)	81.4 (6.8)	79.6 (6.6)	80.8 (6.5)	81.3 (6.9)
Group at Diagnosis, n (%)							
65–69	366 (5.2)	129 (7.1)	86 (4.0)	6 (2.5)	109 (5.4)	30 (4.8)	4 (2.2)
70–74	1,277 (18.3)	411 (22.7)	330 (15.5)	39 (16.5)	384 (19.2)	85 (13.5)	27 (14.8)
75–79	1,753 (25.1)	519 (28.7)	457 (21.5)	50 (21.2)	521 (26.0)	156 (24.8)	49 (26.9)
80–84	1,768 (25.3)	438 (24.2)	550 (25.8)	61 (25.8)	517 (25.8)	159 (25.3)	42 (23.1)
85+	1,832 (26.2)	313 (17.3)	705 (33.1)	80 (33.9)	473 (23.6)	199 (31.6)	60 (33.0)
Race, n (%)							
White	6,573 (94.0)	1,688 (93.3)	2,006 (94.3)	212 (89.80)	1,880 (93.8)	613 (97.5)	168 (92.3)
Black	361 (5.2)	108 (6.0)	101 (4.7)	23 (9.7)	105 (5.2)	13 (2.1)	10 (5.5)
Others	62 (0.9)	14 (0.8)	21 (1.0)	1 (0.4)	19 (0.9)	3 (0.5)	4 (2.2)
Income, mean (SD)	10,602.6 (3,322.7)	10,771.0 (3,277.0)	10,686.5 (3,502.6)	10,611.8 (3,035.0)	10,329.7 (3,195.9)	10,715.4 (3,237.0)	10,521.0 (3,466.5)
Stage, n (%)							
In situ	363 (5.2)	0 (0.00)	144 (6.8)	4 (1.7)	215 (10.7)	0 (0.0)	0 (0.0)
Localized	2,454 (35.1)	407 (22.5)	659 (31.0)	65 (27.5)	1,159 (57.8)	158 (25.1)	0 (0.0)
Regional	1,855 (26.5)	408 (22.5)	859 (40.4)	82 (34.7)	425 (21.2)	81 (12.9)	0 (0.0)
Distant	1,634 (23.4)	710 (39.2)	297 (14.0)	52 (22.0)	108 (5.4)	289 (45.9)	178 (97.8)
Unstaged	690 (9.9)	285 (15.7)	169 (7.9)	33 (14.0)	97 (4.8)	101 (16.1)	4 (2.2)
Method of Diagnosis, n (%)							
Microscopic confirmation	6,358 (90.9)	1,441 (79.6)	2,039 (95.8)	229 (97.0)	1,958 (97.7)	551 (87.6)	133 (73.1)
Laboratory test/ mater study	61 (0.9)	0 (0.0)	3 (0.1)	0 (0.00)	0 (0.0)	37 (5.9)	21 (11.5)
Direct visualization	27 (0.4)	9 (0.5)	14 (0.7)	0 (0.0)	4 (0.2)	0 (0.0)	0 (0.0)
Radiography/ other imaging techniques	321 (4.6)	259 (14.3)	37 (1.7)	3 (1.3)	13 (0.6)	9 (1.4)	0 (0.0)
Clinical diagnosis	111 (1.6)	54 (3.0)	15 (0.7)	3 (1.3)	13 (0.6)	10 (1.6)	16 (8.8)
Unknown	118 (1.7)	47 (2.6)	20 (0.9)	1 (0.4)	16 (0.8)	22 (3.5)	12 (6.6)
Incidence Rates[b]							
Our study population	–	332.4	392.0	43.2	440.7	115.4[d] (33.3)[d]	–
SEER[c]	–	299.6	307.2	45.2	449.7	87.4[e]	54.0[e]

[a] Including male breast cancer

[b] 100,000 person years

[c] Age and gender standardized to our PACE population

[d] Chronic lymphocytic leukemia is included in lymphoma but not in leukemia by our definition based on WHO definition

[e] Chronic lymphocytic leukemia is included in leukemia but not in lymphoma by SEER definition

period. The mean age of the study population was 79 years, 83% were women, 95% were white, 5% were black and 1% was of other race.

Table 1 shows demographic and clinical characteristics of cancer cases identified by the registry within the cohort. In general, the proportion of men and older patients was higher in cancer cases than in the entire cohort. The overall completeness of stage and diagnosis confirmation in the registry was greater than 90% and 98% respectively, and cancers were confirmed microscopically for 91% of cases. The estimated incidence of cancers identified by the registry within the cohort was similar to the age-specific SEER-reported incidence [13].

Specificity, sensitivity, and PPV claims data-based definitions

The number of cases, specificity, sensitivity and PPV of four claims data-based definitions are shown in Table 2. All definitions had very high specificity (greater than 98%), whereas sensitivity varied from 40% to 90%. Differences in specificity were minimal among Definitions 1 to 3, but sensitivity differed considerably among definitions (highest in Definition 4, and lowest in Definition 1). Very small differences in specificity affected PPVs considerably, which was most extreme for leukemia that also had the lowest prevalence among the six cancers in the study population. Definition 3 had very high specificity yet had relatively high sensitivity, and was used for the subsequent analyses. All subsequent analyses were limited to Definition 3. When we calculated sensitivity, specificity, and PPV as a function of age at cohort entry, we observed slight to moderate variability in these measures among age groups; however, no systematic trend was observed across six cancers.

Table 3 shows sensitivity, specificity, and PPV for incident cancers as a function of various cancer-free periods, ranging from 0 to 36 months. Specificity, sensitivity, and PPV were expected to improve as we lengthen the period since fewer prevalent or recurrent cancers would be misclassified as incident in the study period. Although all these measures were greatly improved by changing the required cancer-free period from 0 to 6 months, no meaningful improvement was achieved by further lengthening the period beyond 6 months.

Accuracy of diagnosis dates

The diagnosis date derived from Medicare data tended to come later than the date recorded in the registry data except for breast cancer. The median difference in days is close to

Table 2 Sensitivity, specificity, and positive predictive value of claims-based definitions of incident cancers

		#Cases	Sensitivity (%)	Specificity (%)	PPV (%)
Lung	Definition 1	1,344	56.35	99.79	75.89
	Definition 2	2,088	76.19	99.54	66.04
	Definition 3	2,235	80.06	99.49	64.83
	Definition 4	3,472	86.69	98.78	45.19
Colorectal	Definition 1	2,017	67.25	99.62	70.95
	Definition 2	2,464	80.36	99.51	69.40
	Definition 3	2,799	83.98	99.35	63.84
	Definition 4	4,179	88.02	98.51	44.82
Stomach	Definition 1	276	69.92	99.93	59.78
	Definition 2	345	81.36	99.90	55.65
	Definition 3	383	84.32	99.88	51.96
	Definition 4	602	89.41	99.75	35.05
Breast	Definition 1	1,150	46.91	99.84	81.74
	Definition 2	2,065	78.89	99.62	76.56
	Definition 3	2,232	83.03	99.56	74.55
	Definition 4	3,483	87.23	98.65	50.19
Lymphoma	Definition 1	564	55.17	99.86	61.52
	Definition 2	799	79.81	99.81	62.83
	Definition 3	926	83.31	99.74	56.59
	Definition 4	1,607	88.71	99.33	34.72
Leukemia	Definition 1	185	41.76	99.93	41.08
	Definition 2	220	52.20	99.92	43.18
	Definition 3	297	61.54	99.88	37.71
	Definition 4	712	73.63	99.63	18.82

Definition 1: Combination of diagnosis and procedures on the same day or within the same hospitalization

Definition 2: Two Diagnoses of specific cancer within 2 months

Definition 3: Definition 1 or Definition 2

Definition 4: One diagnosis of cancer

Table 3 Sensitivity, specificity, and positive predictive value of Definition 3 as a function of the required cancer-free periods (0 to 36 months) to distinguish incident cases from recurrent cases

	Total Population	0 month	6 months	12 months	18 months	24 months	30 months	36 months
		260,457	157,310	137,474	122,797	112,243	100,546	92,051
Lung	Sensitivity	77.15	80.06	81.37	82.11	83.14	83.25	83.81
	Specificity	99.36	99.49	99.52	99.53	99.53	99.56	99.57
	PPV	59.88	64.83	65.73	66.04	66.19	66.69	66.92
Colorectal	Sensitivity	81.18	83.98	84.34	84.75	85.55	85.79	86.16
	Specificity	99.22	99.35	99.38	99.38	99.39	99.40	99.40
	PPV	59.51	63.84	64.84	65.30	66.13	66.46	66.52
Stomach	Sensitivity	81.47	84.32	83.33	83.85	84.21	84.62	84.62
	Specificity	99.87	99.88	99.89	99.90	99.90	99.90	99.90
	PPV	48.27	51.96	53.68	55.90	55.17	57.14	57.62
Breast	Sensitivity	79.26	83.03	83.86	84.92	85.86	86.67	87.30
	Specificity	99.44	99.56	99.56	99.56	99.56	99.55	99.56
	PPV	69.99	74.55	74.77	75.26	75.23	74.85	75.30
Lymphoma	Sensitivity	79.58	83.31	84.10	84.73	85.01	85.49	85.99
	Specificity	99.66	99.74	99.76	99.76	99.77	99.77	99.77
	PPV	48.58	56.59	58.26	57.86	58.74	58.72	58.93
Leukemia	Sensitivity	57.43	61.54	63.29	64.75	66.67	67.86	68.00
	Specificity	99.84	99.88	99.89	99.89	99.89	99.89	99.89
	PPV	28.52	37.71	39.68	40.00	41.58	41.76	40.48

PPV, positive predictive value

zero (0 to 2 depending on cancer type), the mean differences were greater than 10 days (12–22 days) except for breast cancer (–1 day). Across different cancers, 21 to 46 % of the cases had their diagnoses recorded on the same day, and another 21 to 48% had their diagnoses recorded within ±7 days. The differences in the diagnosis dates were within ±14 days for 74.1% to 88.0%, and within ±60 days for 85.7% to 97.0% of the cases depending on cancer type.

Impact of residual misclassification of cancer as an outcome on pharmacoepidemiologic studies

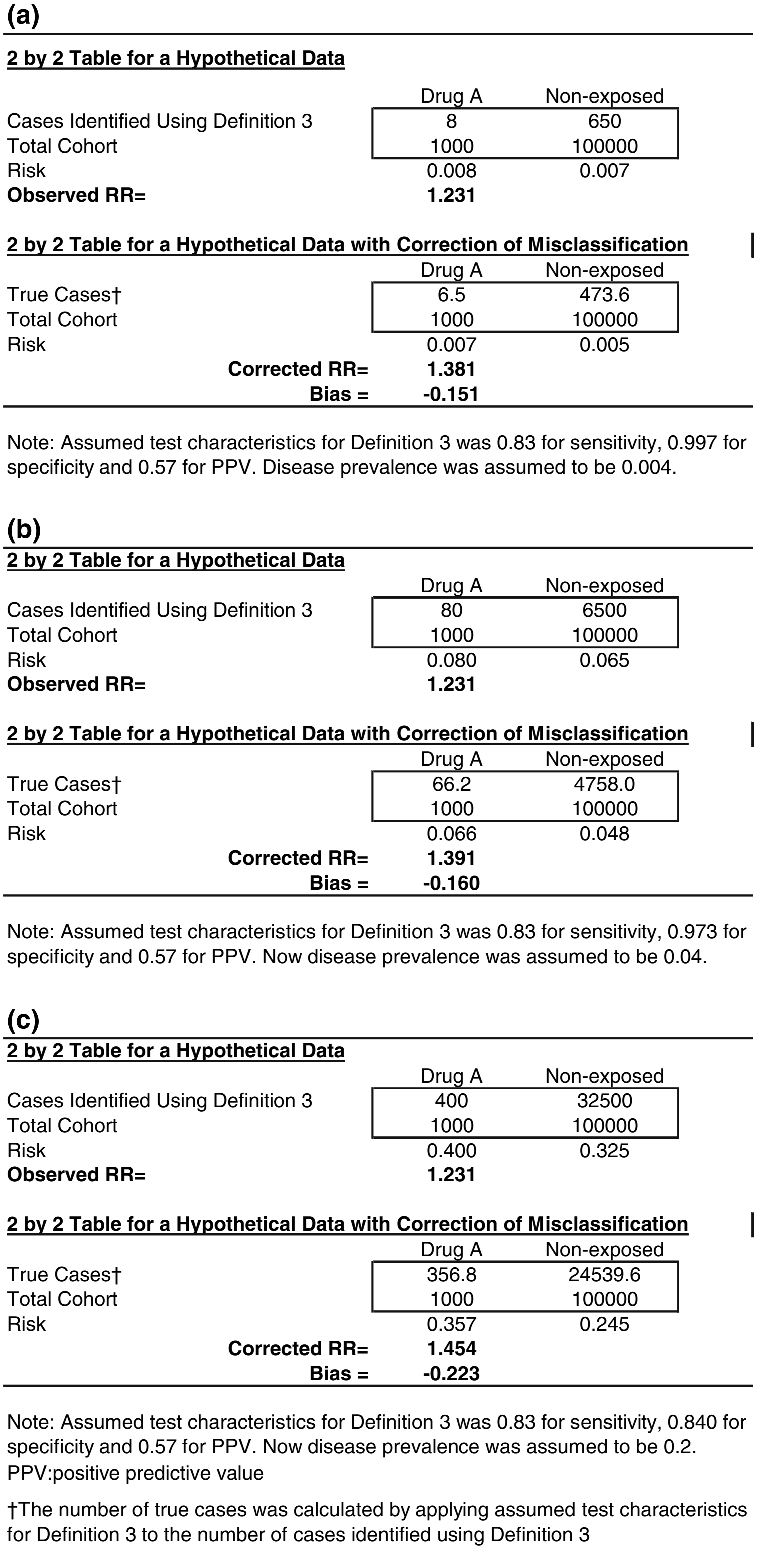
We created data for a hypothetical cohort study evaluating the effect of Drug A on the risk of lymphoma (Table 4a–c). In these data, incident lymphomas were identified using claims data and the effect of Drug A on lymphoma was estimated. We then calculated the expected number of cases and a corrected risk ratio (RR) using estimated sensitivity, specificity, and PPV [14] and disease prevalence in the current study assuming non-differential disease misclassification. (Table 4a) This resulted in a small bias towards the null, which was calculated by subtracting the observed risk ratio (RR), from the corrected RR. Next, we decreased specificity by increasing the disease prevalence but kept the PPV constant to illustrate the impact of specificity and PPV on bias. (Table 4b–c) Although the low PPV observed in our study was a concern, there was not substantial bias when the disease prevalence was low and specificity was very high (Table 4a) However, as specificity decreased, the impact of the same low PPV (0.57) on bias became greater (Table 4b–c)

Discussion

We assessed agreement between Medicare claims-identified cancer cases and registry identified cancer cases, and calculated sensitivity, specificity, and PPV for four Medicare claim-based definitions of cancers using the population-based cancer registry cases as the gold standard. Our study showed that incident hematologic malignancies and solid tumors could be identified using claims data with very high specificity but relatively low sensitivity. We showed that the agreement in the cancer diagnosis dates was reasonably good between claims and registry data.

We also observed that possible misclassification of prevalent cases as incident cases improved most significantly by requiring a 6-month cancer-free period in the claims data but longer periods did not impact sensitivity, specificity and PPV. These findings supports the rationale of having a 6-month required cancer-free period when studying the association between medication use or other risk factors and the incidence of cancer using health care utilization databases. By lengthening the required cancer-free period, we expected that the specificity and PPV would improve as the number of prevalent and recurrent cases

Table 4 The impact of cancer outcome misclassification on rate ratio estimates in a hypothetical cohort study assessing the effect of Drug A on lymphoma using a claims-based Definition. (a) Using sensitivity and specificity of the definition 3 estimated in thestudy. (b) Using the same sensitivity and PPV but smaller specificity and larger disease prevalance. (c) Using the same sensitiviy and PPV but much smaller specificity and larger disease prevalance



(a)

2 by 2 Table for a Hypothetical Data

	Drug A	Non-exposed
Cases Identified Using Definition 3	8	650
Total Cohort	1000	100000
Risk	0.008	0.007
Observed RR=	**1.231**	

2 by 2 Table for a Hypothetical Data with Correction of Misclassification

	Drug A	Non-exposed
True Cases†	6.5	473.6
Total Cohort	1000	100000
Risk	0.007	0.005
Corrected RR=	**1.381**	
Bias =	**-0.151**	

Note: Assumed test characteristics for Definition 3 was 0.83 for sensitivity, 0.997 for specificity and 0.57 for PPV. Disease prevalence was assumed to be 0.004.

(b)

2 by 2 Table for a Hypothetical Data

	Drug A	Non-exposed
Cases Identified Using Definition 3	80	6500
Total Cohort	1000	100000
Risk	0.080	0.065
Observed RR=	**1.231**	

2 by 2 Table for a Hypothetical Data with Correction of Misclassification

	Drug A	Non-exposed
True Cases†	66.2	4758.0
Total Cohort	1000	100000
Risk	0.066	0.048
Corrected RR=	**1.391**	
Bias =	**-0.160**	

Note: Assumed test characteristics for Definition 3 was 0.83 for sensitivity, 0.973 for specificity and 0.57 for PPV. Now disease prevalence was assumed to be 0.04.

(c)

2 by 2 Table for a Hypothetical Data

	Drug A	Non-exposed
Cases Identified Using Definition 3	400	32500
Total Cohort	1000	100000
Risk	0.400	0.325
Observed RR=	**1.231**	

2 by 2 Table for a Hypothetical Data with Correction of Misclassification

	Drug A	Non-exposed
True Cases†	356.8	24539.6
Total Cohort	1000	100000
Risk	0.357	0.245
Corrected RR=	**1.454**	
Bias =	**-0.223**	

Note: Assumed test characteristics for Definition 3 was 0.83 for sensitivity, 0.840 for specificity and 0.57 for PPV. Now disease prevalence was assumed to be 0.2.
PPV:positive predictive value

†The number of true cases was calculated by applying assumed test characteristics for Definition 3 to the number of cases identified using Definition 3

misclassified as incident cancer cases decrease. In addition to these improvements, we also observed some increase in sensitivity, because the size of the cohort (the denominator) decreased with longer cohort membership requirement and it decreased faster than the number of cases identified by a claims data-based definition. This fast decrease in sample size reflected high turnover of subjects in the PACE program and similar to what is observed in HMOs.

Cooper et al., [4] examined the sensitivity of Medicare claims for six common cancers (breast, colorectal, endometrial, lung, prostate, and pancreatic cancers) using the SEER-Medicare database. They found that the sensitivity of a corresponding cancer diagnoses or a cancer-specific procedure coded in outpatient or hospital files for lung, colorectal, and breast cancers was 80% to 90%. Freeman et al. [8] examined the sensitivity, specificity, and PPV of a prediction model for incident breast cancer in the SEER–Medicare database using logistic regression. Using their optimal cut-point, the sensitivity, specificity, and PPV were 90%, 99.9% and 70%, respectively. We found that the sensitivity, specificity, and PPV were 83%, 99.6%, and 75%, respectively, using the preferred definition for breast cancer, which is comparable to Freeman's findings. Our results for lung, colorectal and breast cancers are consistent with these previous studies.

A few limitations of our study should be noted. Our findings of low sensitivity and PPV especially in hematologic cancers may be partly explained by incomplete ascertainment of cases by the PA cancer registry [15–19]. Although, cancer registries make every effort to capture all cases, the case ascertainment may not be 100% and cases captured by claims data may enhance the cancer surveillance [16, 18]. The impact of incomplete case ascertainment can be large especially in rare cancers and could underestimate their PPVs to a large extent. We illustrated this in a figure to show how sensitivity, specificity and PPV of the claims-based definition changes depending on the degree of case ascertainment of the registry (alloyed gold standard) assuming that the missed cases by the registry were captured by our claims-based definition (Fig. 2). When we assumed that the registry captures only 80% of the true cases [16, 18], the PPV for lymphoma using our preferred definition will be 74% (compared to 56% if the registry captures 100% of the true cases), whereas the estimates for sensitivity and specificity are not affected substantially.

To calculate sensitivity, specificity, and PPV with our claims data-based definitions, we linked the subjects in the cohort sampled from Medicare/PACE with the registry data using person-specific identifiers. Because the subjects in our cohort were a subset of the entire Medicare population in PA, e.g., those also enrolled in the PACE program, whereas the registry data include all cases in PA, we could not assess how successful the linkage was. It was impossible to assess whether we had non-linked registry cases because of poor linkage or because they arose outside of the cohort. However, we had several personal identifiers available for linkage, and the successful linkage was indirectly supported by the findings of our study; incidence in the cohort was similar to that in the SEER database, and sensitivity, specificity and PPV in some of the cancers were compatible with previous studies. Any insufficiency in linkage would likely diminish specificity and PPV.

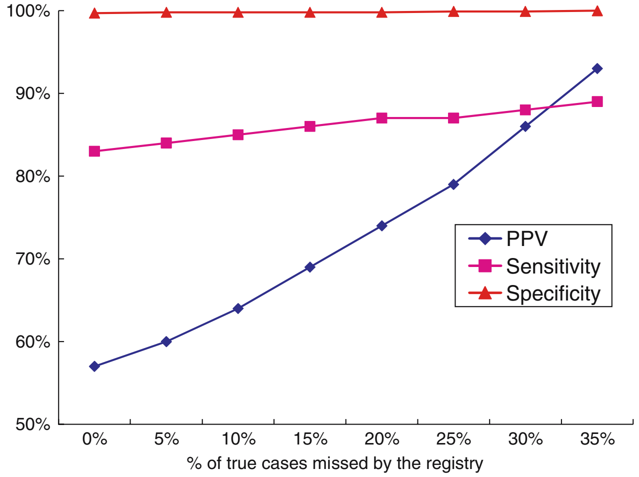


Fig. 2 Estimated sensitivity, specificity, and positive predictive values (PPV) for a claims-based definition (Definition 3) of lymphoma vary depending on the degree of incomplete case ascertainment by the registry in the main analyses, we assumed that there are no cases missed by the registry (% of true cases missed by the registry = 0) to calculate sensitivity, specificity and PPV of the claims-based definitions. When we vary the assumption of the percentage of cases missed by the registry, as the percentage of cases missed by the registry increases, the estimated values of these test characteristics measures increased. The change was the largest in PPV but smaller in sensitivity and much smaller in specificity

We conclude that claims data can identify incident hematologic malignancies and solid tumors with high specificity and but with relatively low to moderate sensitivity and PPVs. Within the cases identified by both the registry and the claims-based definition, the agreement in the first dates of cancer diagnosis was sufficient. Our claims-based definition resulted in relatively small bias in our example of a typical pharmacoepidemiologic study with drug A possibly causing lymphoma. However, the impact of bias due to misclassification and thus the usefulness of claims-based cancer definitions as cancer outcome markers in etiologic studies need to be assessed for each study setting.

Acknowledgments The authors thank Jeffrey Peppercorn, MD, MPH, Assistant Professor of Medicine, Division of Hematology/Oncology, University of North Carolina at Chapel Hill for useful comments on the health care utilization data-based definitions for incident cancers and the draft version of this work. The authors also thank Helen Mogun, MS for her contributions to linking the data of this work. *Financial Support:* This project was support by a grant from the Engalitcheff Arthritis Outcomes Initiatives, Arthritis Foundation. Dr. Schneeweiss received support from the National Institute on Aging (RO1-AG021950, RO1- AG023178) and the Agency for Healthcare Research and Quality (2-RO1-HS10881), Department of Health and Human Services, Rockville, MD, and is principal investigator of the BWH DEcIDE Research Center funded by the Agency for Healthcare Research and Quality.

Reference

1. Brown SL, Greene MH, Gershon SK, Edwards ET, Braun MM (2002) Tumor necrosis factor antagonist therapy and lymphoma development: twenty-six cases reported to the Food and Drug Administration. Arthritis Rheum 46(12):3151–3158
2. Setoguchi S, Solomon DH, Weinblatt ME et al (2006) Tumor necrosis factor alpha antagonist use and cancer in patients with rheumatoid arthritis. Arthritis Rheum 54(9):2757–2764
3. Warren JL, Feuer E, Potosky AL, Riley GF, Lynch CF (1999) Use of Medicare hospital and physician data to assess breast cancer incidence.[comment]. Medical Care 37(5):445–456
4. Cooper GS, Yuan Z, Stange KC, Dennis LK, Amini SB, Rimm AA (1999) The sensitivity of Medicare claims data for case ascertainment of six common cancers.[comment]. Medical Care 37(5):436–444
5. Solin LJ, Legorreta A, Schultz DJ, Levin HA, Zatz S, Goodman RL (1994) Analysis of a claims database for the identification of patients with carcinoma of the breast. J Med Syst 18(1):23–32
6. Solin LJ, MacPherson S, Schultz DJ, Hanchak NA (1997) Evaluation of an algorithm to identify women with carcinoma of the breast. J Med Syst 21(3):189–199
7. Leung KM, Hasan AG, Rees KS, Parker RG, Legorreta AP (1999) Patients with newly diagnosed carcinoma of the breast: validation of a claim-based identification algorithm. J Clin Epidemiol 52(1):57–64
8. Freeman JL, Zhang D, Freeman DH, Goodwin JS (2000) An approach to identifying incident breast cancer cases using Medicare claims data. J Clin Epidemiol 53(6):605–614
9. Penberthy L, McClish D, Manning C, Retchin S, Smith T (2005) The added value of claims for cancer surveillance: results of varying case definitions. Med Care 43(7):705–712
10. The North American Association of Central Cancer Registries, Inc. Official website (2004) (Accessed September, 9, 2005, at http://www.naaccr.org/.)
11. Fletcher RW, Fletcher SW (2005) Clinical Epidemiology: The Essentials. 4th edn. Lippincott Williams & Wilkins (ed)
12. Rothman KJ, Greenland S (1998) Modern Epidemiology. 2nd edn. Lippincott Williams & Wilkins
13. SEER Fast Stats (Accessed January 15, 2005, 2005, at http://seer.cancer.gov.)
14. Brenner H, Gefeller O (1993) Use of the positive predictive value to correct for disease misclassification in epidemiologic studies. Am J Epidemiol 138(11):1007–1015
15. Fanning J, Gangestad A, Andrews SJ (2000) National cancer data base/surveillance epidemiology and end results: potential insensitive-measure bias. Gynecol Oncol 77(3):450–453
16. Penberthy L, McClish D, Pugh A, Smith W, Manning C, Retchin S (2003) Using hospital discharge files to enhance cancer surveillance. Am J Epidemiol 158(1):27–34
17. Stang A, Glynn RJ, Gann PH, Taylor JO, Hennekens CH (1999) Cancer occurrence in the elderly: agreement between three major data sources. Ann Epidemiol 9(1):60–67
18. Wang PS, Walker AM, Tsuang MT, Orav EJ, Levin R, Avorn J (2001) Finding incident breast cancer cases through US claims data and a state cancer registry. Cancer Causes & Control 12(3):257–265
19. McClish DK, Penberthy L, Whittemore M et al (1997) Ability of Medicare claims data and cancer registries to identify cancer cases and treatment. Am J Epidemiol 145(3):227–233