HHS Public Access

Author manuscript

Med Care. Author manuscript; available in PMC 2019 December 01.

Published in final edited form as:

Med Care. 2018 December ; 56(12): e83–e89. doi:10.1097/MLR.0000000000000875.

Leveraging Linkage of Cohort Studies with Administrative Claims Data to Identify Individuals with Cancer

Mackenzie R. Bronson, BA[1], **Nirav S. Kapadia, MD, MS**[1,2], **Andrea M. Austin, PhD**[1], **Qianfei Wang, MS**[1], **Diane Feskanich, ScD**[3], **Julie P.W. Bynum, MD, MPH**[1], **Francine Grodstein, ScD**[3,*], and **Anna N.A. Tosteson, ScD**[1,2,*]

[1]The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine at Dartmouth, Lebanon, NH

[2]Norris Cotton Cancer Center, Geisel School of Medicine at Dartmouth, Lebanon, NH

[3]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA

Abstract

Background—In an effort to overcome quality and cost constraints inherent in population-based research, diverse data sources are increasingly being combined. In this paper, we describe the performance of a Medicare claims-based incident cancer identification algorithm in comparison to observational cohort data from the Nurses' Health Study (NHS).

Methods—NHS-Medicare linked participants' claims data were analyzed using four versions of a cancer identification algorithm across three cancer sites (breast, colorectal, and lung). The algorithms evaluated included an update of the original Setoguchi et al. algorithm, and three other versions that differed in the data used for prevalent cancer exclusions.

Results—The algorithm that yielded the highest positive predictive value (0.52–0.82) and kappa statistic (0.62–0.87) in identifying incident cancer cases utilized both Medicare claims and observational cohort data (NHS) to remove prevalent cases. The algorithm that only used NHS data to inform the removal of prevalent cancer cases performed nearly equivalently in statistical performance (PPV: 0.50–0.79; Kappa: 0.61–0.85) while the version that used only claims to inform the removal of prevalent cancer cases performed substantially worse (PPV: 0.42–0.60; Kappa: 0.54–0.70), in comparison to the dual data source informed algorithm.

Conclusions—Our findings suggest claims-based algorithms identify incident cancer with variable reliability when measured against an observational cohort study reference standard. Self-reported baseline information available in cohort studies is more effective in removing prevalent cancer cases than are claims data algorithms. Use of claims-based algorithms should be tailored to the research question at hand and the nature of available observational cohort data.

Corresponding Author: Mackenzie Bronson, One Medical Center Drive, WTRB Level 1, Lebanon, NH 03756, mackenzie.bronson@dartmouth.edu, 603.653.0800.

*Drs. Grodstein and Tosteson both served as senior authors

Introduction

Population-based research in cancer generally depends on large cohorts and data obtained from diverse sources, which may include prospective cohorts, administrative claims data, and state or national cancer registries. Each source individually may be limited by the quality of the information or the cost and effort in obtaining it. Yet, combining these data sources could maximize the quality and efficiency of data collection in addressing public health questions. For example, claims data may be utilized to replace expensive follow-up of cohort participants to identify incident cancers. However, prior research has not examined how identification of cancer cases may compare in administrative claims data versus epidemiologic cohort studies, and whether the two sources could be combined in simple ways that improve identification.

In this manuscript, we first updated a claims-based algorithm developed to identify incident cancers (Setoguchi et al. 2008[1]), and compared its performance against cancer diagnoses identified in Nurses' Health Study (NHS) participants as the reference standard. Second, we tested combinations of the claims algorithm and the NHS data in an effort to improve our ability to identify new cancer diagnoses. In particular, claims data is limited in its ability to exclude prevalent diagnoses, whereas many cohorts may themselves collect information on prevalent diseases at study initiation. Third, we examined the performance of a modified claims algorithm in identifying prevalent cancers compared to NHS. Fourth, we compared the date of incident cancer diagnosis in the claims and NHS data.

Methods

Overall, we searched the literature for published algorithms to identify incident cancers using claims data, and then both updated the algorithm as well as explored how the algorithm might be improved by the simple incorporation of NHS cohort data for excluding prevalent cancer cases at baseline. We evaluated the performance of these algorithms using NHS data as a reference standard. We focused on three cancers with differing incidence and mortality: breast, colorectal, and lung cancers. Details of our methods are described below. This study received institutional review board approval.

Reference Standard, Nurses' Health Study (NHS)

The NHS cohort was initiated in 1976, when 121,700 female registered nurses, age 30–55 years, living in 11 U.S. states, responded to a mailed questionnaire about their health and lifestyle. Questionnaires are sent biennially and inquire about diagnoses, including specific cancers, during the previous two years; questionnaires are mailed repeatedly throughout each two-year cycle until a reply is elicited.

An NHS participant was initially eligible for the Medicare claims linkage if her social security number was available (80%), if she did not "opt out" of the Medicare linkage, and she responded to an NHS questionnaire following the observation window (or died during the observation window) to ensure loss to follow-up did not harm thorough reporting or identification of cases.

Medicare Claims Data

Medicare claims were restricted to female enrollees aged 65 and older before July 1 2007 with continuous enrollment in Medicare Parts A and B over the two-year study period from 2008–2009. Characteristics of NHS-Medicare linked subjects were obtained from Medicare Beneficiary files, and their utilization was based on MEDPAR and Carrier Files.

Of the NHS reported cases (breast, colorectal, and lung cancers) with diagnosis dates during the study period from 2008–2009, a large majority (71.8%) were confirmed by medical records and another 20.6% were re-confirmed by nurse-participant in the follow-up letter or call (and the remainder were confirmed by participant only or death record). Additionally, for some NHS cancer cases, the date of diagnosis assigned may be slightly late (e.g., if the cancer case is identified by the NHS at death rather than at exact time of diagnosis, which is particularly common for rapidly fatal cancers, such as lung); thus, we allowed NHS confirmation to occur in the succeeding twelve months (through 2010) – this accounted for only 9% of cancer cases, but helped reduce the inappropriate appearance of “false positive” cases found in claims data (i.e., cases identified in Medicare in 2008–2009 but which were found in NHS in 2010).

Claims Algorithm for Incident and Prevalent Cancers

A literature review of published, claims-based methods to identify incident cancer cases yielded several algorithms[2–15], each with varying purposes and designs. We narrowed this list to algorithms that relied upon U.S. administrative claims data, that addressed multiple cancer sites, and that presented clear, reproducible methods. Among the options, the Medicare claims-based algorithm by Setoguchi, et al. was the only algorithm that met all of these criteria. In that publication, and in the present work, incident cancer was defined as a cancer that was first diagnosed within the study period while prevalent cancer was defined as a cancer that was diagnosed prior to the study period (we did not try to identify recurrent cancers).

The Setoguchi algorithm consists of six rules for identifying incident cancers using Medicare claims (Supplementary Table 1). Each of the six identification rules utilize specific International Classification of Disease (ICD) codes for diagnosis[16] and/or Current Procedural Terminology (CPT)[17] codes, which are part of the Medicare billing process. The ICD-9 codes are specific to each type of cancer while the CPT codes encompass the various types of cancer treatment. The six identification rules range from criteria such as cancer diagnosis (e.g., 2 cancer ICD-9 codes within 60 days) to cancer treatment (e.g., chemotherapy CPT codes plus cancer ICD codes). To address whether the identified cancer was incident versus prevalent, the algorithm uses claims data from six-months prior to the start of the study period to remove prevalent cancer cases from the eligibility for incident cancer identification. Setoguchi recommended the six-month window, after review of various durations up to 36 months, without improvement of algorithm performance.

In our application of this algorithm, we reproduced the six identification rules with some modifications. We used a two-year observation window, which identified incident cancers from January 1, 2008 - December 31, 2009, instead of the four-year period used by

Setoguchi. We did this for practical reasons: 2008–9 represents a time period for which NHS has had time to complete all documentation and data management of new cancer cases and deaths (which can take several years), and this time period corresponds to the NHS follow-up cycles which start each even year. To the original Setoguchi algorithm we also both updated the original list of CPT codes used in their 2007 publication, and added pathology codes that reflected clinical procedures routinely employed in the reporting of cancer after tissue sampling (e.g., cytopathology review). See Supplementary Table 2 for details.

In assessing the Setoguchi algorithm's application for cancer detection in the context of an epidemiologic cohort study (NHS), we considered the updated Setoguchi algorithm (#1). To examine how claims and cohort data might be combined to exclude prevalent cancers prior to the study period, we developed two further versions, which varied how prevalent cancers were identified and excluded using claims plus cohort data (#2 and #3). In addition, one version (#4) focused on identification of any cancer without distinguishing between prevalent and incident cancers. These four versions are detailed below and depicted in Figure 1:

#1. Updated incident cancer identification, utilizing 6-month claims data window to exclude cancer prior to study period: This algorithm is updated and a slight modification (as described above) of the Setoguchi algorithm.

#2. Incident cancer identification, utilizing NHS data to exclude prevalent cancer at any point prior to study period: This algorithm leveraged the longer follow-up period available in NHS, and removed prevalent cancer diagnoses identified by NHS from 1976–2007. This algorithm did not use claims data for searching prevalent cases.

#3. Incident cancer identification, utilizing 6-month window in claims data and NHS to exclude prevalent cancer: This algorithm combined #1 and #2, and removed prevalent cancers identified either in claims data from the previous 6 months or in NHS data from 1976 through 2007. Thus, this algorithm utilizes all of the information on prevalent cancers available on cohort participants in both data sources.

#4. Prevalent cancer identification, utilizing claims only: We also considered an algorithm in which prevalent cancers were not excluded, for situations in which distinguishing between prevalent and incident cancers is not necessary.

Validation of Date of Diagnosis

In a separate exercise to compare date of incident cancer diagnosis in claims versus NHS data, we used the algorithm that excluded prevalent cancers based on claims and NHS data (#3) to establish our pool of incident cancers. Since NHS identifies only month and year of diagnosis, we also used only month and year for claims-based incident cancer diagnoses. The assigned claims-based diagnosis date varied according to the type of service that was performed and recorded in the claims (for details see Supplementary Table 2).

Statistical Analysis

Summary statistics were calculated for each of the four algorithms comparing the claims-based detection of breast, colorectal, and lung cancers to cases reported in NHS. Standard

summary statistics included sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). To characterize algorithm performance across these summarize statistics, NHS data was the reference standard. In addition, the kappa statistic was calculated to assess the agreement between claims-detected cancer diagnoses and NHS-identified cancer diagnoses.

To characterize the performance of each of the six individual rules within the algorithms for identifying cancer diagnoses in the claims data (see Supplementary Table 1), we plotted the sensitivity (true positive rate) against one minus specificity (false positive rate) for each identification rule for each cancer site using the algorithm which applied both NHS and claims data for excluding prevalent cancers.

Results

There were 41,809 female NHS participants for whom claims data linkage was available in the relevant time period. Participant characteristics are displayed in Table 1. In 2008, women ranged from 65 to 85 years of age, were predominately non-Hispanic white, distributed throughout the primary U.S. census regions (although the largest number lived in the Northeast), similar to the overall NHS study population. During the two-year study period, NHS identified 872 incident cancer cases (454 breast, 152 colorectal, and 266 lung cancer cases).

We applied the four versions of the algorithm (see Figure 1) across breast, colorectal, and lung cancer cases, and calculated a range of performance measures (Table 2). The population sample varied somewhat between algorithms due to the differences in exclusions applied. In the basic algorithm (#1), sensitivity ranged from 93% for breast, to 79% for colorectal and 85% for lung cancers; in addition, PPV was relatively low (44% for breast, 42% for colorectal, 60% for lung cancer). In contrast, specificity was very high across all cancer sites, as was negative predictive value (approximately 99% for both measures for each cancer). Overall, the kappa statistic was also relatively low (0.59 for breast, 0.54 for colorectal, 0.70 for lung cancer).

When we excluded prevalent cases using all NHS data (#2), sensitivity remained similar to algorithm #1, suggesting that studies could rely on whichever source was simpler or less expensive to obtain, if sensitivity was of the highest importance. However, using NHS data for the exclusion increased PPV somewhat, to 79% for breast, 50% for colorectal, and 64% for lung cancers, while specificity and NPV remained approximately 99% across cancer sites. In addition, the kappa statistic improved when using NHS data to exclude prevalent cancers (0.85 for breast, 0.61 for colorectal, 0.73 for lung cancers).

Finally, for the algorithm that used both claims and NHS data to exclude prevalent cancers (#3), results for sensitivity, specificity, PPV, NPV and kappa were all nearly identical to the algorithm that utilized only NHS data (#2).

The algorithm which searched for all cancers (#4), performed best across all summary statistics. For breast cancer, sensitivity, specificity, and NPV were all approximately 99%, with PPV of 96% and kappa statistic of 0.97. The algorithm also performed well for

colorectal and lung cancers, with sensitivity of 93% for colorectal and 92% for lung, specificity 99% for both, PPV 77% for colorectal and 80% for lung cancer, NPV 99% for both, and kappa statistics of 0.84 and 0.85, respectively.

Since the original Setoguchi algorithm involves six different "rules" (see Supplementary Table 1), we also examined the operating characteristics of each rule (Figure 2) as applied to algorithm #3. We found that detection rates of both true positives (sensitivity) and false positives (1-specificity) varied as a function of rule and cancer site. For instance, rates of true and false positive rates of cancer detection based on use of biopsy claims varied little, ranging from 0.92–0.97 and 0.59–0.63, respectively, depending upon the cancer. In contrast, the operating characteristics of surgical claim-based identification of cancer resulted in broader variation, with true positive rates ranging from 0.30 for lung cancer to 0.78 for colon cancer and false positives ranging from 0.10 for lung cancer to 0.50 for colon cancer. Notably, use of chemotherapy claims resulted in very low accuracy of breast cancer identification, with a lower rate of true than false positive detection.

When we examined date of diagnosis in the two data sources, using only the true positive cases from the algorithm #3, the month and year were within three months of each other for claims versus NHS data in 86.7% (350/404) of breast cancer cases, 82.3% (93/113) of colorectal cancer cases, and 86.9% (179/206) of lung cancer cases. Moreover, for cases with a difference in diagnosis dates of more than 3 months, the median difference between the claims-derived and NHS-derived month and year of diagnosis was just 6 months (70 with claims coming first, 31 with NHS coming first) across the three cancer sites.

Discussion

We assessed a published claims-based algorithm to identify cancers, and compared claims-identified cases with cases identified via contacting participants every two years from the NHS cohort. In addition, since the ability to exclude prevalent cases is a major challenge in claims data, we also explored whether combining claims and "baseline" cohort data might improve algorithm performance. Motivation for undertaking this study was to delineate the challenges and opportunities inherent in augmenting observational data from epidemiological cohort studies with administrative claims data that provide detailed information on use of health care services to identify individuals with incident or prevalent cancer. Indeed, our ability to identify incident cancer cases using claims data was best (highest PPV and kappa) when all information was utilized in removing prevalent cases. Nonetheless, the nearly identical performance characteristics when we did not use any claims data for removing prevalent cases suggests that cohorts with self-reported, baseline information from participants about cancer history may not benefit appreciably from buying or using claims data prior to the study period. Moreover, the modest sensitivity, and particularly low PPV and kappa when using only claims data, for a 6-month window, to exclude prevalent cases underscores some of the limitations of identifying incident cancer cases using claims data alone. These issues were largely a result of a fair number of false positive (prevalent, not incident) cases found using claims, although extending the 6-month window in claims data to 12 months in our study, or to 36 months in the Setoguchi publication, had no meaningful impact on algorithm performance. Nonetheless, if there were

means of “re-confirming” case status (e.g., only following up on participants with claims-based cancer diagnoses rather than the entire cohort) then claims could yield cost-savings in large population-based cohorts compared to repeatedly contacting all participants.

Additionally, algorithm #4 was successful specifically in identifying true positive cases, although many were diagnosed in the years prior to the study period. The claims data clearly include many women receiving cancer-related health care services and/or health care services labeled with cancer diagnoses for cancers diagnosed years earlier. This may reflect on-going care (e.g., cancer related complication), active surveillance (e.g., biopsy), or could also be due to continued coding of diagnoses in subsequent years. Regardless, in research where both incident and prevalent cancers case identification is useful, the claims data algorithm performed well.

Overall our application of the Setoguchi algorithm to a cohort study confirmed findings from the Setoguchi publication, where claims data were compared to state cancer registry data, but we did find some differences. Namely, our sensitivity for detecting breast and lung cancer were superior to Setoguchi, though our ability to detect colorectal cancer was inferior to Setoguchi’s report (77.4% vs 83.9%). Possible reasons for these dissimilarities include: differences in cancer prevalence between the two studies, the lesser ability of claims to distinguish incident from prevalent cancer diagnoses, and dissimilar participant characteristics (i.e., NHS participants are all female and mostly Caucasian).

The distribution of the various algorithm identification rules shown in Figure 2 demonstrated the tradeoff between sensitivity and specificity. Although some identification rules performed similarly across the cancers, others varied considerably by site. For example, radiation therapy identified few colorectal cancer cases (<11, Supplementary Table 3), due to the overall rarity of its use in colon cancer. Alternatively, since surgery is near-universally used to treat breast cancer, its claims-based detection rate translated into both a high true positive (0.66) and low false positive rate (0.34). Based on their variant performances, identification rules could be simplified, and individually used or not based on both the site of cancer one is interested in identifying and the degree of false positive identification one is willing to accept.

The majority of diagnosis dates were identical between the claims-derived diagnosis date (month and year) and the NHS diagnosis date across the three cancer sites. Thus, claims may serve to augment cohort data in this capacity in cases where the date of diagnosis is not available for the participant.

Interpretation of the algorithm results should consider properties of the data sources we utilized. Our goal was to compare a claims-based algorithm to repeated contact with participants, generally considered the standard in population-based epidemiologic research. The two-year claims observation window performed comparably to Setoguchi’s four-year window, suggesting it is plausible to use fewer years of claims data for commensurate results. Additionally, we had varying sample sizes for the different cancer sites, and thus results for colorectal cancer, the least common site here, may be less accurate than for breast cancer.

Our evaluation of the claims algorithms applied to the long-standing NHS helped to highlight several strengths and weaknesses in the use of claims-based algorithms for identifying breast, colorectal or lung cancer in the context of epidemiologic studies (Table 3). While the highest performance in identifying incident cancer diagnoses is achieved with use of both data sources at study baseline, the use of baseline exclusions from participant reports resulted in very similar accuracy; thus cohorts may not need to purchase claims data prior to the baseline date of their study if they have baseline cancer reports available from participants, which is likely common.

Although our study used Medicare claims data exclusively, the algorithms should be broadly applicable to other claims-based data. For instance, the databases from all-payer claims, Medicaid, or specific commercial plans could be leveraged. Such claims data could also provide specific advantages over Medicare, including identifying cancers in those under age 65 years or enabling longer periods of observation at certain ages (potentially providing improved exclusion of prevalent cases, whereas a limitation of Medicare is the lack of information on diagnoses prior to age 65 years or prior to enrollment).

Our study underscores the ability of claims data to identify all cancers when distinguishing between prevalent and incident diagnoses is unimportant. This will be useful in situations where explicit information regarding timing of diagnosis is not needed, such as some genetic research or when establishing survivorship studies, or is available from the cohort study. Our results demonstrate that researchers should consider their research question and data availability when determining how best to use claims-based cancer identification algorithms to advance our understanding of cancer etiology, care, and outcomes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Cancer Institute (grant number UM1CA186107).

The authors above certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest, or any other conflict of interest related to the manuscript other than the above-mentioned funding.

References

1. Setoguchi S, Solomon DH, Glynn RJ, et al. 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):561–9. [PubMed: 17447148]
2. Baldi I, Vicari P, Di Cuonzo D, et al. A high positive predictive value algorithm using hospital administrative data identified incident cancer cases. J Clin Epidemiol. 2008; 61(4):373–9. [PubMed: 18313562]
3. Butler A, Olshan A, Kshirsagar A, et al. Cancer incidence among US Medicare ESRD patients receiving hemodialysis, 1996–2009. Am J Kidney Dis. 2015; 65(5):763–772. [PubMed: 25662835]
4. Couris CM, Polazzi S, Olive F, et al. Breast cancer incidence using administrative data: correction with sensitivity and specificity. J Clin Epidemiol. 2009; 62(6):660–6. [PubMed: 19070463]

5. Fenton J, Onega T, Zhu W, et al. Validation of a Medicare claims-based algorithm for identifying breast cancer detected at screening mammography. Med Care. 2016; 54(3):e15–e22. [PubMed: 23929404]

6. Foley K, Shi N, Girvan A, et al. Claims data algorithms for identifying incident breast cancer (BC) cases and cancer disease stage: a critical review of the literature. Value in Heath. 2013:A1–A298.

7. Freeman JL, Zhang D, Freeman DH, Goodwin JS. An approach to identifying incident breast cancer cases using Medicare claims data. J Clin Epidemiol. 2000; 53(6):605–14. [PubMed: 10880779]

8. Ganry O, Taleb A, Peng J, et al. Evaluation of an algorithm to identify incident breast cancer cases using DRGs data. Eur J Cancer Prev. 2003; 12(4):295–9. [PubMed: 12883382]

9. Gold H, Do H. Evaluation of three algorithms to identify incident breast cancer in Medicare claims data. Health Serv Res. 2007; 42(5):2056–2069. [PubMed: 17850533]

10. Leung KM, Hasan AG, Rees KS, et al. Patients with newly diagnosed carcinoma of the breast: validation of a claim-based identification algorithm. J Clin Epidemiol. 1999; 52(1):57–64. [PubMed: 9973074]

11. Nattinger A, Laud P, Bajorunaite R, et al. An algorithm for the use of Medicare claims data to identify women with incident breast cancer. Health Serv Res. 2004; 36(6 Pt 1):1733–1750.

12. Quantin C, Benzenine E, Hagi M, et al. Estimation of national colorectal-cancer incidences using claims databases. J Cancer Epidemiol. 2012

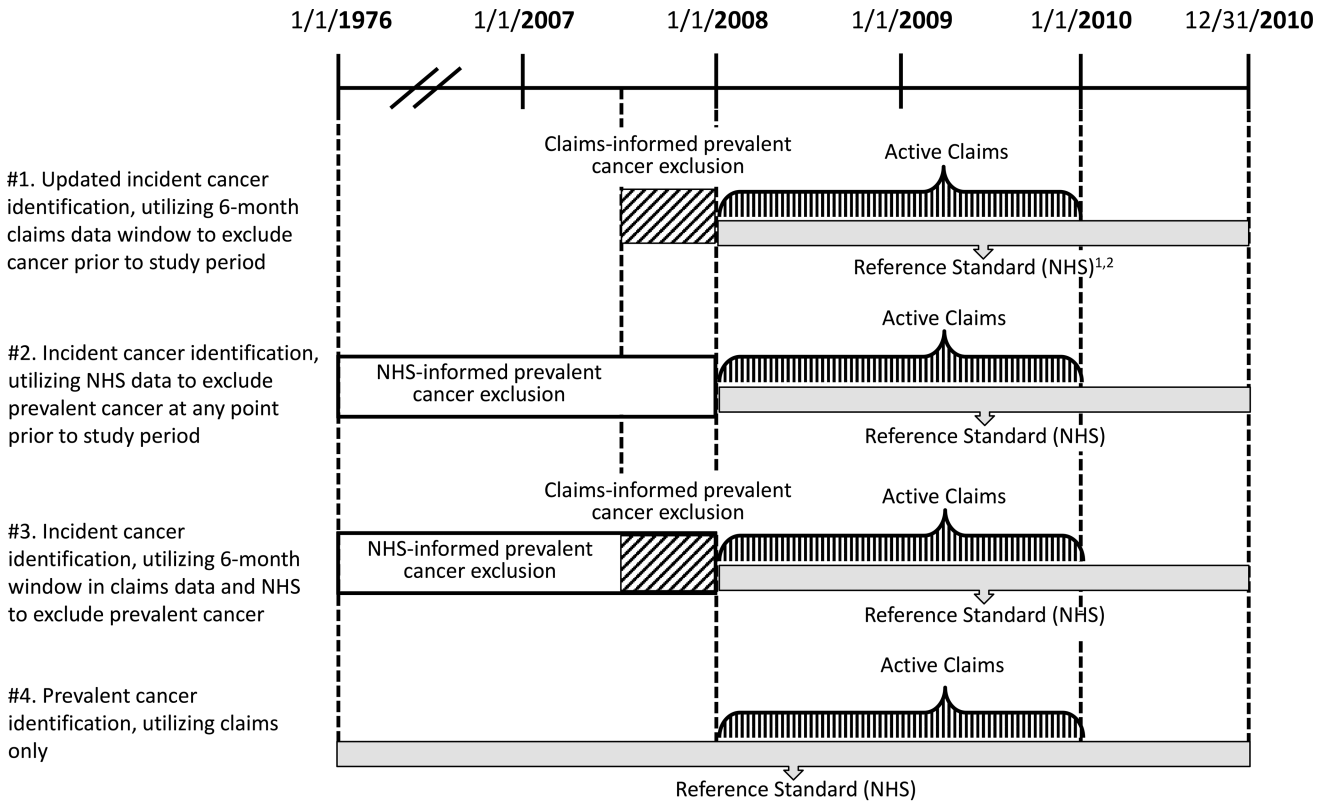
13. Wang P, Walker A, Tsuang M, et al. Finding incident breast cancer cases through US claims data and a state cancer registry. Cancer Causes Control. 2001; 12(3):257–265. [PubMed: 11405331]

14. Warren JL, Feuer E, Potosky AL, et al. Use of Medicare hospital and physician data to assess breast cancer incidence. Med Care. 1999; 37(5):445–56. [PubMed: 10335747]

15. Yuen E, Louis D, Cisbani L, et al. Using administrative data to identify and stage breast cancer cases: implications for assessment quality of care. Tumori. 2011; 97(2):428–35. [PubMed: 21989429]

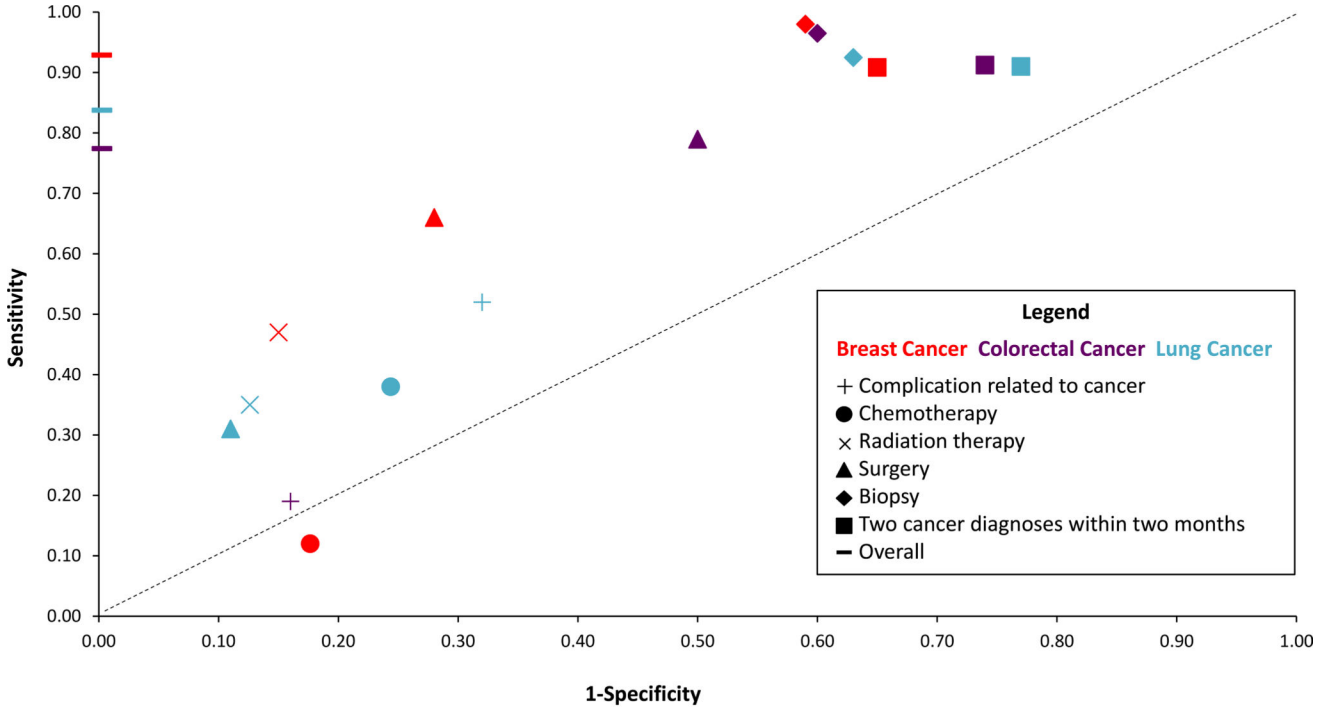
16. World Health Organization. [Accessed March 11, 2017] Classifications of Diseases (ICD). http://www.who.int/classifications/icd/en/

17. American Medical Association. [Accessed March 11, 2017] CPT. https://www.ama-assn.org/practice-management/cpt



[1] Reference Standard: Nurses' Health Study (NHS) self-report of cancer with or without medical record review for confirmation.
[2] The twelve-month reference standard window beyond the active claims observation window is to accommodate the possibility of imprecise diagnosis dates of the NHS questionnaire.

Figure 1.
Four algorithms to identify cancer cases using Medicare claims and Nurses' Health Study data.



[1]Data points with less than eleven observations are removed from the figure.

Figure 2.

Performance of individual rules for identifying cancer in claims data, by cancer site, using algorithm #3.

Table 1

Characteristics of NHS-Medicare linked cohort in 2008*.

Characteristic	NHS-Medicare linked 2008–2009	
	N	%
Total	41,809	
Age		
65–69	9,492	22.70
70–74	11,481	27.46
75–79	9,752	23.33
80+	11,084	26.51
Race		
White	40,574	97.05
Black	523	1.25
Asian/Pacific Islander	280	0.67
Other	432	1.03
Ethnicity		
Hispanic	309	0.74
Zip level median household income (2010)		
Mean	$65,060	-
Std Dev	$24,493	-
25%tile	$46,942	-
Median (IQR)	$59,710	-
75%tile	$77,283	-
Census Region		
Midwest	6,702	16.03
Northeast	19,593	46.86
South	10,256	24.53
West	5,258	12.58
Total Cancer Cases from NHS Reference Standard**		
Breast Cancer	2456	-
Colorectal Cancer	446	-
Lung Cancer	539	-

*All female, Medicare FFS eligible since 07/01/2007 or earlier.

**From algorithm version 4 (true positives plus false negative cases).

Table 2

Comparison of algorithm versions by cancer site.

Algorithm	Total N	True Positive	False Positive	False Negative	True Negative	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	Kappa
Breast Cancer										
1. Updated incident cancer identification, utilizing 6-month claims data window to exclude cancer prior to study period	39,435	416	540	31	38,448	0.9306	0.9861	0.4351	0.9992	0.5866
2. Incident cancer identification, utilizing NHS data to exclude prevalent cancer at any point prior to study period	37,216	422	114	32	36,648	0.9295	0.9969	0.7873	0.9991	0.8506
3. Incident cancer identification, utilizing 6-month window in claims data and NHS to exclude prevalent cancer	37,092	404	88	31	36,569	0.9287	0.9976	0.8211	0.9992	0.8700
4. Prevalent cancer identification, utilizing claims only	41,809	2,424	114	32	39,239	0.9870	0.9971	0.9551	0.9992	0.9689
Colorectal Cancer										
1. Updated incident cancer identification, utilizing 6-month claims data window to exclude cancer prior to study period	41,370	121	169	33	41,047	0.7857	0.9959	0.4172	0.9992	0.5428
2. Incident cancer identification, utilizing NHS data to exclude prevalent cancer at any point prior to study period	40,977	119	121	33	40,704	0.7829	0.9970	0.4958	0.9992	0.6054
3. Incident cancer identification, utilizing 6-month window in claims data and NHS to exclude prevalent cancer	40,888	113	105	33	40,637	0.7740	0.9974	0.5183	0.9992	0.6193
4. Prevalent cancer identification, utilizing claims only	41,809	413	121	33	41,242	0.9260	0.9971	0.7734	0.9992	0.8410
Lung Cancer										
1. Updated incident cancer identification, utilizing 6-month claims data window to exclude cancer prior to study period	41,475	219	146	40	41,070	0.8456	0.9965	0.6000	0.9990	0.6997
2. Incident cancer identification, utilizing NHS data to exclude prevalent cancer at any point prior to study period	41,380	224	124	42	40,990	0.8421	0.9970	0.6437	0.9990	0.7277
3. Incident cancer identification, utilizing 6-month window in claims data and NHS to exclude prevalent cancer	41,291	206	104	40	40,941	0.8374	0.9975	0.6645	0.9990	0.7393
4. Prevalent cancer identification, utilizing claims only	41,809	497	124	42	41,146	0.9221	0.9970	0.8003	0.9990	0.8549

Table 3

Algorithm versions' strengths, weaknesses, and suggested uses.

Algorithm	Strengths	Weaknesses	Qualitative Summary of Findings
1. Updated incident cancer identification, utilizing 6-month claims data window to exclude cancer prior to study period	Requires only claims data with a narrow additional time window (six months) for excluding prevalent cases	Limited time period (six months) for excluding prevalent cancer cases	Lowest specificity and comparatively low PPV and Kappa points to limitations in using claims data alone for identification of incident cancer cases.
2. Incident cancer identification, utilizing NHS data to exclude prevalent cancer at any point prior to study period	Claims not used to exclude prevalent cancer cases.	Moderate PPV and kappa, especially for colorectal cancer	Can be applied when data on cancer history are obtained at cohort inception to ensure only incident cases are identified through claims
3. Incident cancer identification, utilizing 6-month window in claims data and NHS to exclude prevalent cancer	Makes full use of both data sources	Very close performance characteristics to Algorithm #2.	Higher specificity results in small improvement in PPV and kappa. Use of both data sources minimizes false positive incidenct cancer diagnoses with minimal change in sensitivity.
4. Prevalent cancer identification, utilizing claims only	Only requires claims from a two-year observation window to identify those who have ever had cancer	Cannot distinguish incident from prevalent cases	High sensitivity, specificity, PPV, NPV, kappa for identifying ever cancer diagnoses. Useful in when date of diagnosis is not required (eg, studies of genetic factors, or of early-life risk factors for adult cancers) or if diagnosis date is available from other sources.