HHS Public Access

Author manuscript

Pharmacoepidemiol Drug Saf. Author manuscript; available in PMC 2017 April 01.

Published in final edited form as:

Pharmacoepidemiol Drug Saf. 2016 April ; 25(4): 472–475. doi:10.1002/pds.3953.

Identification of smoking using Medicare data- A validation study of claims-based algorithms

Rishi J. Desai[1], **Daniel H. Solomon**[1,2], **Nancy Shadick**[2], **Christine Iannaccone**[2], and **Seoyoung C. Kim**[1,2]

[1]Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital & Harvard Medical School, Boston, MA, USA

[2]Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital & Harvard Medical School, Boston, MA, USA

Abstract

Purpose—This study examined the accuracy of claims-based algorithms to identify smoking against self-reported smoking data.

Methods—Medicare patients enrolled in the Brigham and Women's Hospital Rheumatoid Arthritis Sequential Study (BRASS) were identified. For each patient, self-reported smoking status was extracted from BRASS and the date of this measurement was defined as the index-date. Two algorithms identified smoking in Medicare claims; 1) only using diagnoses and procedure codes, and 2) using anti-smoking prescriptions in addition to diagnoses and procedure codes. Both algorithms were implemented first only using 365-days pre-index claims and then using all available pre-index claims. Considering self-reported smoking status as the gold standard, we calculated specificity, sensitivity, positive predictive value (PPV), negative predictive value (NPV), and area under the curve (AUC).

Results—A total of 128 patients were included in this study, of which 48% reported smoking. The algorithm only using diagnosis and procedure codes had the lowest sensitivity (9.8%, 95% CI 2.4%–17.3%), NPV (54.9%, 95% CI 46.1%–63.9%), and AUC (0.55, 95% CI 0.51–0.59) when applied in the period of 365 days pre-index. Incorporating pharmacy claims and using all available pre-index information improved the sensitivity (27.9%, 95% CI 16.6%–39.1%), NPV (60.4%, 95% CI 51.3%–69.5%), and AUC (0.64, 95% CI 0.58–0.70). The specificity and PPV was 100% for all the algorithms tested.

Conclusion—Claims-based algorithms can identify smokers with limited sensitivity but very high specificity. In the absence of other reliable means, use of a claims-based algorithm to identify smoking could be cautiously considered in observational studies.

Correspondence: Rishi J Desai, MS, PhD, Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 1620 Tremont Street, Suite 3030-R, Boston, MA 02120, USA, Phone: 617-278-0932 | Fax: 617-232-8602, rdesai2@partners.org.

Conflict of interest:

Dr. Solomon's work on this project was funded by the NIH (K24-AR055989). He also receives salary support through research support to his hospital from Amgen, Pfizer, AstraZeneca, Genentech, and CORRONA. Dr. Kim is supported by the NIH grant K23 AR059677. She receives research grants from AstraZeneca, Lilly, Pfizer and Genentech.

Keywords

Claims-based algorithm; smoking; validation

Introduction

Large healthcare utilization claims databases from around the world are being increasingly used in comparative effectiveness and safety studies of drug treatments.[1–3] In order to draw reliable inferences, it is important to use valid algorithms to identify clinical conditions and patient characteristics of interest, which may be used as inclusion criteria, exclusion criteria, or confounders in these studies. However, as the primary purpose of healthcare utilization claims is billing, certain behavioral characteristics of the patients are usually not directly recorded in these data sources, making their ascertainment especially challenging.

One such important patient characteristic is tobacco use or smoking, which is a risk factor for a large number of chronic diseases[4] and therefore is of considerable interest while conducting observational studies for confounding control purposes through restriction or statistical adjustment. Smoking status is not directly available in claims data sources used for pharmacoepidemiolgy research.[5] However, claims data often contain information on indirect indicators for smoking, including diagnosis of tobacco use disorder, records for counseling visits for smoking, and anti-smoking prescription medication use, which can serve as a proxy measure for smoking. These indirect indicators have been used to identify smokers in observational studies conducted using claims data from various insurance programs including the Department of Veteran's Affairs,[6] Medicaid,[7] as well as Medicare.[8] However, limited information exists regarding the accuracy of identifying smoking based solely on these alternative indicators in claims data.

Therefore, we sought to examine the accuracy of several claims based approaches for the identification of patients with smoking using patient-reported smoking data from a cohort of Medicare-insured rheumatoid arthritis (RA) patients enrolled in the Brigham and Women's Hospital Rheumatoid Arthritis Sequential Study (BRASS).

Methods

The BRASS registry is a single-center, prospective and observational cohort of 1,350 patients with rheumatologist-verified diagnosis of RA. For the subjects enrolled in this registry, data on patient reported items including demographics, medication use, lifestyle factors including smoking status and alcohol use, and quality of life scales, as well as physician reported items such as extra-articular manifestations, and medication changes are collected during annual follow-up visits. For this study, we identified patients from BRASS who were also enrolled in Medicare between 2006 and 2010, and linked their data from these two sources. Of these subjects, we further identified those with at least one BRASS visit with valid self-reported smoking status after 365 days continuous enrollment in Medicare parts A, B, and D. The BRASS visit date with a validly recorded self-reported smoking status was defined as the index-date.

Two algorithms (Table 1) were implemented using Medicare claims data to identify smoking for these patients in two periods, 1) 365 days pre-index date, and 2) all available information pre-index date. Briefly, the first algorithm only used diagnoses codes and procedure codes potentially related to smoking, while the second algorithm used pharmacy claims in addition to diagnoses and procedure codes to define smoking. Considering self-reported smoking status (as ever or never) as the gold standard, we calculated specificity, sensitivity, positive predictive value (PPV), negative predictive value (NPV), and area under the curve (AUC) for both claims-based algorithms. Asymptotic 95% confidence intervals were reported for each of these measures.

Results

We identified 350 patients who were enrolled in both BRASS and Medicare and had at least one valid self-reported smoking status indicator recorded during their BRASS visit. The final sample consisted of 128 patients with at least 365 days of continuous enrollment in Medicare parts A, B, and D prior to their eligible BRASS visit date. The cohort represented 89% White, 6% Black, and 5% patients of other races. The mean [standard deviation (SD)] age was 69 (10) years and 88% of the sample were female. In this cohort, 48% of the patients reported ever-smoking.

Table 2 summarizes the performance of various claims-based approaches for identification of smoking. The first algorithm, which only used diagnoses codes and procedure codes, had the lowest sensitivity (9.8%, 95% CI 2.4%–17.3%), NPV (54.9%, 95% CI 46.1%–63.9%), and AUC (0.55, 95% CI 0.51–0.59) when applied in the period of 365 days pre-index. Incorporating 365-days pre-index pharmacy claims in the second algorithm resulted in improvement of sensitivity (26.2%, 95% CI 15.2%–37.3%), NPV (59.8%, 95% CI 50.1%–69.8%), and AUC (0.58, 95% CI 0.53–0.63). Using all available pre-index claims resulted in greater sensitivity, NPV and AUC for both algorithms compared with only using 365-days prior claims. The best performing approach for identifying smoking was using all available diagnosis codes, procedure codes, and prescription claims pre-index (algorithm 2), with a sensitivity of 27.9% (95% CI 16.6%–39.1%), NPV of 60.4% (95% CI 51.3%–69.5%), and AUC of 0.64 (95% CI 0.58–0.70). Notably, the specificity and PPV was 100% across the four approaches tested.

In the approach with the highest sensitivity, patients accurately identified as smokers by the claims-based algorithm had a longer duration of self-reported smoking compared with smokers that were not identified by the algorithm [mean (standard deviation (SD)) 35 (16) years versus 24 (15) years, $p<0.05$ (t-test)].

Discussion

In this validation study, we assessed the accuracy of several approaches to identify smoking status from claims data against self-reported smoking status using a cohort of Medicare-insured RA patients enrolled in a prospective registry. Overall, we observed that the claims based algorithms had low sensitivities and NPVs, but combining data from medical and pharmacy claims as well as using all available pre-index information to determine smoking

increased the sensitivity and NPVs. Conversely, all these algorithms had excellent specificities and PPVs.

Two previous studies have examined the validity of administrative data-based algorithms to identify smoking in different patient populations. Wiley et al.[9] used ICD-9 codes from the electronic medical records (EMRs) to identify smoking and compared it with a 'gold-standard' smoking definition derived from chart review of full-text of the EMRs. This study reported sensitivity of 32% and specificity of 100% for the ICD-9 code based algorithm. In a second study, Kim et al.[10] used ICD-9 codes from administrative claims combined with EMR chart note data from the department of Veteran's Affairs to define smoking and reported 39% sensitivity and 98% specificity for this algorithm against a 'gold-standard' smoking definition derived from full text chart-review.

The sensitivity of the most comprehensive algorithm in our study (27%) was somewhat lower compared with the two aforementioned studies, which may be explained by two important differences between this study and the prior investigations. First, we defined smoking status solely based on administrative claims and did not use EMR information to reflect the fact that a majority of the claims data sources available in the US do not contain linked EMR information. Second, we had patients' self-reported smoking data available to us, while the previous studies relied on full-text chart reviews to create a gold-standard measure for smoking, which may have lower sensitivity than self-reports.

This study provides important implications for pharmacoepidemiologists who wish to measure smoking status in their investigations using claims data. First, the claims-based smoking algorithm could be reliably used to create a cohort of smokers owing to its high PPV. However, low sensitivity of this algorithm also means that such approach could result in exclusion of a large number of true smokers, limiting the sample size of the planned study. Second, if the aim of a study is to compare health outcomes between smokers and non-smokers, the claims-based algorithm may have limited value due to low NPV. The misclassification of smokers into non-smoker group due to low NPV could bias the results towards the null in such investigations. Third, if the aim is to control for confounding by smoking status, one must be alert to the possibility of residual confounding even after adjustment due to low sensitivity of this algorithm in circumstances where the prevalence of smoking is drastically different between the exposed and the reference groups.[11]

Our study has several strengths. It is the first study evaluating the performance of claims based algorithms combining procedure codes and prescription claims with diagnosis codes to identify smoking in a commonly used US administrative claims data-source. Detailed information available in BRASS allowed us to compare the smoking history between smokers successfully identified using claims-based algorithm and missed by the algorithm, providing important insights into the performance of this algorithm. Our study also has some limitations including a small sample size and use of a patient cohort with diagnosis of a chronic condition (RA) potentially inflating the sensitivity of the algorithms as they are more likely to have used healthcare services in the past. The prevalence for ever-smoking among US adults reported by the Centers for Disease Control and Prevention is 42.5%,[12] which is somewhat lower compared with 48% observed in our cohort. Therefore, our study

may have marginally overestimated the PPV and underestimated the NPV of claims-based algorithms.

In conclusion, Medicare claims data can estimate smoking with limited sensitivity but very high specificity. In the absence of other reliable means, use of a claims-based algorithm to identify smoking could be cautiously considered in studies focusing on smokers and in studies where confounding control by smoking status is desirable.

Acknowledgments

Funding information:

This research was not funded through any external institution.

References

1. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epi. Apr; 2005 58(4):323–337.
2. Hoffmann F. Review on use of German health insurance medication claims data for epidemiological research. Pharmacoepidemiol Drug Saf. May; 2009 18(5):349–356. [PubMed: 19235771]
3. Martin-Latry K, Begaud B. Pharmacoepidemiological research using French reimbursement databases: yes we can! Pharmacoepidemiol Drug Saf. Mar; 2010 19(3):256–265. [PubMed: 20128015]
4. Carter BD, Abnet CC, Feskanich D, et al. Smoking and mortality--beyond established causes. NEJM. Feb 12; 2015 372(7):631–640. [PubMed: 25671255]
5. Brookhart MA, Sturmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. Med Care. Jun; 2010 48(6 Suppl):S114–120. [PubMed: 20473199]
6. Kurtz SM, Lau E, Ong KL, et al. Infection risk for primary and revision instrumented lumbar spine fusion in the Medicare population. J Neurosurg Spine. Oct; 2012 17(4):342–347. [PubMed: 22920611]
7. Desai RJ, Huybrechts KF, Hernandez-Diaz S, et al. Exposure to prescription opioid analgesics in utero and risk of neonatal abstinence syndrome: population based cohort study. BMJ. 2015; 350:h2102. [PubMed: 25975601]
8. Chen D, Restrepo MI, Fine MJ, et al. Observational study of inhaled corticosteroids on outcomes for COPD patients with pneumonia. Am J Respir Crit Care Med. Aug 1; 2011 184(3):312–316. [PubMed: 21512168]
9. Wiley LK, Shah A, Xu H, Bush WS. ICD-9 tobacco use codes are effective identifiers of smoking status. JAMIA. Jul-Aug;2013 20(4):652–658. [PubMed: 23396545]
10. Kim HM, Smith EG, Stano CM, et al. Validation of key behaviourally based mental health diagnoses in administrative data: suicide attempt, alcohol abuse, illicit drug abuse and tobacco use. BMC Health Serv Res. 2012; 12:18. [PubMed: 22270080]
11. Schonberger RB, Gilbertsen T, Dai F. The problem of controlling for imperfectly measured confounders on dissimilar populations: a database simulation study. J Cardiothorac Vasc Anesth. Apr; 2014 28(2):247–254. [PubMed: 23962461]
12. Centers for Disease Control and Prevention (CDC). Tobacco use among adults--United States, 2005. MMWR Morb Mortal Wkly Rep. 2006 Oct 27; 55(42):1145–8. [PubMed: 17065979]

Key points

- Limited information exists regarding the accuracy of identifying smoking in claims data. We used Medicare claims data linked with rheumatoid arthritis registry data from the Brigham and Women's Hospital to validate claims based smoking algorithms against self-reported smoking status.
- Claims-based algorithms can identify smokers with limited sensitivity but very high specificity. In the absence of other reliable measures of smoking, use of a claims-based algorithm to identify smoking could be cautiously considered in studies focusing on smokers and in studies where confounding control by smoking status is desirable.

Table 1

Algorithms to identify smoking from Medicare claims data

Algorithm 1 (Medical claims based only)	Presence of at least one of the following codes on at least one inpatient or outpatient medical claim ICD 9 codes 305.1→ Tobacco use disorder 649.0x → Tobacco use complicating pregnancy 989.84 → Toxic effect of tobacco V15.82 → Personal history of tobacco use CPT codes 99406, 99407, G0436, G0437, G9016 → Smoking counseling visits S9453 → Smoking cessation classes S4995 → Smoking cessation gum G9276, G9458 → Documented tobacco user advised to quit 1034F → Current smoker 4004F, 4001F → Screened for tobacco use and received an intervention
Algorithm 2 (Medical claims plus pharmacy claims based)	Presence of one of the ICD-9 or CPT codes listed in Algorithm 1 on at least one inpatient or outpatient medical claim OR dispensing of at least one nicotine or varenicline prescription

Abbreviations: CPT- Current Procedure Terminology, ICD- International Classification of Diseases.

Table 2

Performance of Medicare-claims based algorithms in identifying smokers using self-reported smoking status as the gold-standard

Measure*	Algorithm 1 (Medical claims only)		Algorithm 2 (Medical and Pharmacy claims)	
	365 days pre-index	All available information pre-index	365 days pre-index	All available information pre-index
Sensitivity	9.8% (2.4%–17.3%)	26.2% (15.2%–37.3%)	16.4% (7.1%–25.7%)	27.9% (16.6%–39.1%)
Specificity	100% (94.6%–100%)	100% (94.6%–100%)	100% (94.6%–100%)	100% (94.6%–100%)
PPV	100% (54.1%–100%)	100% (79.4%–100%)	100% (69.1%–100%)	100% (80.4%–100%)
NPV	54.9% (46.1%–63.9%)	59.8% (50.1%–69.8%)	56.8% (47.8%–65.7%)	60.4% (51.3%–69.5%)
AUC	0.55 (0.51–0.59)	0.63 (0.58–0.69)	0.58 (0.53–0.63)	0.64 (0.58–0.70)

Abbreviations: AUC- Area under the curve, NPV- Negative predictive value, PPV- Positive predictive value

*Total sample size of 128 patients (67 self-reported non-smokers and 61 smokers) with enrollment in BRASS and Medicare (parts A, B, and D). Numbers in the bracket are 95% confidence intervals.