Annals of Epidemiology 24 (2014) 666–672

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Original article

Limited validity of diagnosis codes in Medicare claims for identifying cancer metastases and inferring stage

Neetu Chawla PhD, MPH [a,b,*], K. Robin Yabroff PhD, MBA [b], Angela Mariotto PhD [c], Timothy S. McNeel BA [d], Deborah Schrag MD, MPH [e], Joan L. Warren PhD [b]

[a] Outcomes Research Branch, Applied Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Rockville, MD
[b] Health Services and Economics Branch, Applied Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Rockville, MD
[c] Data Modeling Branch, Surveillance Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Rockville, MD
[d] Information Management Services, Inc, Calverton, MD
[e] Dana-Farber Cancer Institute, Boston, MA

ARTICLE INFO

Article history:
Received 5 February 2014
Accepted 17 June 2014
Available online 3 July 2014

Keywords:
Cancer
Metastasis
SEER
Registry
Medicare claims
Stage at diagnosis

ABSTRACT

Purpose: Researchers are using diagnosis codes from health claims to identify metastatic disease in cancer patients. The validity of this approach has not been established.
Methods: We used the linked 2005–2007 Surveillance, Epidemiology and End Results (SEER)–Medicare data to assess the validity of metastasis codes at diagnosis from claims compared with stage reported by SEER cancer registries. The cohort included 80,052 incident breast, lung, and colorectal cancer patients aged 65 years and older. Using gold-standard SEER data, we evaluated sensitivity, specificity, positive predictive value, and negative predictive value of claims-based stage, survival by stage classification, and patient factors associated with stage misclassification using multivariable regression.
Results: For patients with a registry report of distant metastatic cancer, the sensitivity, specificity, and positive predictive value of claims never simultaneously exceeded 80% for any cancer: lung (42.7%, 94.8%, and 88.1%), breast (51.0%, 98.3%, and 65.8%), and colorectal (72.8%, 93.8%, and 68.5%). Misclassification of stage from Medicare claims was significantly associated with inaccurate estimates of stage-specific survival ($P < .001$). In adjusted analysis, patients who were older, black, or living in low-income areas were more likely to have their stage misclassified in claims.
Conclusions: Diagnosis codes in Medicare claims have limited validity for inferring cancer stage and metastatic disease.

Published by Elsevier Inc.

Introduction

Researchers are increasingly using administrative claims data to identify cancer metastasis and infer stage at diagnosis and recurrence among cancer patients [1–7]. However, the validity of diagnostic codes for metastasis in claims data has not been well established in population-based data. A prior study examined the accuracy of metastasis codes to infer stage for six common cancers and concluded that Medicare claims have limited utility for defining cancer stage at diagnosis compared with cancer registry data [1]. However, this study only included patients diagnosed between 1984 and 1993, and the accuracy of metastasis coding may have improved since that time. Additionally, this study only examined hospital claims and assessment of physician claims may also be useful in identifying metastases.

More recent studies have suggested that hospital and physician claims may have utility for assessing metastasis, but these studies have been generally restricted to single academic institutions and small samples, limiting their generalizability [2,4]. For instance, Thomas et al. [2] examined claims data to identify lung cancer stage and concluded that the metastases codes were useful for patients seen in an academic institution with private insurance, but should be used with caution on a broader basis. No recent studies have included large population-based cohorts or evaluated the possible implications of misclassification on stage-specific survival.

In this study, we assessed the accuracy of metastasis codes from health claims using the population-based linked Surveillance, Epidemiology and End Results (SEER)–Medicare data. Specifically, we compared the validity of metastasis codes from Medicare claims in the period after cancer diagnosis with SEER historic stage as the

* Corresponding author. Outcomes Research Branch, Applied Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, 9609 Medical Center Drive, Rockville, MD 20850. Tel.: +1 240 276 6896; fax: +1 240 276 7883.
E-mail address: neetu.chawla@nih.gov (N. Chawla).

1047-2797/$ – see front matter Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.annepidem.2014.06.099

gold standard for three common cancers in the United States—breast, colorectal, and lung cancers. We also assessed the impact of misclassification on stage-specific survival and whether inferring stage from Medicare claims results in systematic stage misclassification for any patient groups. Findings from this study have direct relevance for research using Medicare and other administrative claims data to identify metastasis and infer stage at the time of diagnosis.

Methods

Data sources

We used the linked SEER–Medicare data for this study. The SEER population-based registries include nine states (California, Connecticut, Hawaii, Iowa, Kentucky, Louisiana, New Jersey, New Mexico, and Utah) and six metropolitan areas (Atlanta, Detroit, Los Angeles, San Francisco–Oakland, San Jose–Monterey, and Seattle), representing approximately 28% of the US population [8]. SEER registries have detailed reporting guidelines and extensive training efforts to instruct registrars on coding of stage [9–11]. SEER registries also engage in quality improvement initiatives and have contractual obligations to meet specific data quality goals and develop methods to prevent and correct errors in the data [12].

The SEER–Medicare linkage was first completed in 1991 and has been updated on a regular basis since that time [13,14]. To link SEER with Medicare data, the registries participating in the SEER program send individual identifiers for all persons in their files which are matched, using a deterministic algorithm, to identifiers contained in Medicare's master enrollment file. For each year's linkage, 93% of persons aged 65 years and older in the SEER files were matched to the Medicare enrollment file. Further detail on the process of matching individuals from SEER data with Medicare records has been described elsewhere [14].

For each patient, the SEER data contain a unique case number, each occurrence of a primary incident cancer, month and year of diagnosis, tumor stage at diagnosis, treatment information, and date and cause of death for patients who have died. The Medicare data include all hospital, physician, and outpatient clinic claims for Medicare covered services from enrollment until death for beneficiaries with fee-for-service coverage among those aged 65 years and older [8].

Sample selection

From the SEER–Medicare data, we selected all patients aged 65 years and older who were diagnosed between January 1, 2005 and December 31, 2007 with breast, lung, or colorectal cancers ($n = 158{,}262$). Individuals were excluded from the cohort for the following reasons: males with breast cancer ($n = 441$); month of diagnosis was unknown ($n = 691$); SEER month of death was unknown ($n = 5$); SEER historic stage was *in situ* or unknown ($n = 19{,}566$); did not have continuous part A/B fee-for-service enrollment from diagnosis month until 2 months after diagnosis month or death ($n = 40{,}806$); and died before or at 2 months after the diagnosis month ($n = 16{,}701$). After exclusions, a total of 80,052 breast, lung, and colorectal cancer patients with local, regional, or distant disease were included in this study.

Measures

Defining stage at diagnosis

We used SEER historic stage as the gold standard because this is most comparable to how metastasis codes in Medicare claims have been used to infer stage. SEER historic stage uses both clinical and pathologic documentation of the extent and spread of disease obtained from the medical record [10,11]. Coding instructions for SEER historic stage follow guidelines established by the North American Association of Central Cancer Registries [10]. In the Medicare claims data, *International Classification of Diseases*, Ninth Revision, Clinical Modification diagnosis codes for secondary malignant neoplasms to specific organs were used to identify metastases, which were defined as regional or distant metastases for each cancer site (See Appendix Table 1). We reviewed hospital inpatient and physician claims from diagnosis month until 2 months afterward and classified patients as having metastases if they had either a single inpatient claim with metastasis code(s) or two physician claims on separate days with metastasis codes, as has been done elsewhere [15,16]. The requirement for 2 days of physician claims with metastases codes helps to eliminate inaccuracies in coding that can occur when "ruling out" metastases as part of diagnostic workup. Patients without any metastases codes in inpatient claims or with metastases codes only in physician claims on a single day were classified as having local disease.

Because patients could have multiple diagnosis codes for metastases in their claims that could result in their classification as having both regional and distant disease at diagnosis, we used a sequential strategy to classify stage as either regional or distant. Within the inpatient and physician claims files, the strategy used: (1) diagnoses codes on each claim to classify that claim; (2) claims for each day to classify the day; and (3) days with claims to classify the patient as having either regional or distant disease. For example, a claim with one regional and two distant metastases codes was classified as a distant claim, a day with one regional claim and two distant claims was classified as a distant day, and a patient with one regional day and two distant days was classified as having distant disease at diagnosis. If there were equal numbers of regional and distant claims or days, individuals were classified as having distant disease. There was a small percentage of individuals (1.3%) who did not have any Medicare claims and were classified as having local disease. Finally, because some researchers are interested in assessing the presence of any metastasis, rather than the extent of the metastases, we also created a summary measure for any metastases (i.e., regional or distant).

As part of a sensitivity analyses, we also evaluated a less stringent definition of Medicare claims–based stage, and only required a single claim with a metastases diagnosis code in either the inpatient or physician files. We also evaluated a 4-month diagnosis window to assess the impact of the evaluation interval on our findings.

Accuracy measures

We calculated the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of the Medicare claims to infer stage, using SEER historic stage as the gold standard. For each patient, we created a variable for whether stage was misclassified by claims. Among those with stage misclassification, we also assessed whether claims would result in an earlier stage classification than registry or a more advanced stage classification than registry.

Evaluation of patient factors associated with stage misclassification

To assess whether the accuracy of Medicare claims to identify metastatic cancer varied by patient characteristics, we examined age at diagnosis (65–69, 70–74, 75–79, and 80+ years), race (white, black, and other/unknown), gender, and median census tract income in 2000 quartiles (lowest: <$34,456; second: $34,456–$45,760; third; $45,671–$61,234; highest: $61,235+). Each patient's Charlson comorbidity score (0, 1, and 2+) was measured by reviewing diagnoses reported on hospital and