NIH Public Access
Author Manuscript
Ann Epidemiol. Author manuscript; available in PMC 2015 September 01.

Published in final edited form as:
Ann Epidemiol. 2014 September ; 24(9): 666–672.e2. doi:10.1016/j.annepidem.2014.06.099.

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

Neetu Chawla, Ph.D., MPH,
Outcomes Research Branch, Applied Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, 9609 Medical Center Drive Rockville, MD 20850. tel: 240-276-6896, fax: 240-276-7883

K. Robin Yabroff, PhD, MBA,
Health Services and Economics Branch, Applied Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute

Angela Mariotto, PhD,
Data Modeling Branch, Surveillance Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute

Timothy S. McNeel, BA,
Information Management Services, Inc

Deborah Schrag, MD, MPH, and
Dana-Farber Cancer Institute, 450 Brookline Avenue, Boston, MA 02215 Office

Joan L. Warren, PhD
Health Services and Economics Branch, Applied Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute

Neetu Chawla: neetu.chawla@nih.gov

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 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 ages 65 and older. Using gold-standard SEER data, we evaluated sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of claims-based stage, survival by stage-classification, and patient factors associated with stage misclassification using multivariable regression.

Correspondence to: Neetu Chawla, neetu.chawla@nih.gov.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Results—For patients with a registry report of distant metastatic cancer, the sensitivity, specificity, and PPV of claims never simultaneously exceeded 80% for any cancer: lung (42.7%, 94.8%, 88.1%), breast (51.0%, 98.3%, 65.8%), and colorectal (72.8%, 93.8%, 68.5%). Misclassification of stage from Medicare claims was significantly associated with inaccurate estimates of stage-specific survival ($p<0.001$). In adjusted analysis, patients who were older, Black, or living in low-income areas were more likely to have their stage misclassified in claims.

Conclusion—Diagnosis codes in Medicare claims have limited validity for inferring cancer stage and metastatic disease.

Keywords

cancer; metastasis; SEER; registry; Medicare claims; stage at diagnosis

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 to 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 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 utilized with caution on a broader basis.[2] 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 following cancer diagnosis with SEER historic stage as the gold standard for three common cancers in the U.S—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.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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 U.S. 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] In order 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 percent of persons age 65 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 that 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 ages 65 and older.[8]

Sample selection

From the SEER-Medicare data, we selected all patients ages 65 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 ≤ 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 pathological documentation of the extent and spread of disease obtained from the medical record.[10,11] Coding instructions for

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

SEER historic stage follow guidelines established by The North American Association of Central Cancer Registries (NAACCR).[10] In the Medicare claims data, ICD-9-CM 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 afterwards 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 work-up. 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 1 regional and 2 distant metastases codes was classified as a distant claim, a day with 1 regional claim and 2 distant claims was classified as a distant day, and a patient with 1 regional day and 2 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, 80+), race (White, Black, 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+).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Each patient's Charlson comorbidity score (0, 1, 2+) was measured by reviewing diagnoses reported on hospital and physician Medicare claims in the year prior to cancer diagnosis. Given that our cohort was composed of cancer patients, we used a version of the Charlson index that 17 excluded cancer diagnoses, as has been done elsewhere.[17]

Data Analysis

For each cancer site, we compared aggregate stage distribution from SEER data with aggregate stage distribution inferred from Medicare claims. At the individual patient level, we calculated the sensitivity, specificity, PPV, and NPV of stage inferred from the patient's Medicare claims relative to the gold-standard SEER data for all cancers.

To explore the implications of stage misclassification on survival, we compared stage-specific overall survival for subsets of patients where SEER and Medicare claims-based stage classification agreed and disagreed. Survival was calculated from the first day of the month of diagnosis until death or December 31, 2007, the date of data censoring. We used p-values generated from the log-rank test for homogeneity of survival curves to evaluate of the impact of stage misclassification on survival and calculated 95% confidence intervals. Although we examined survival for all stages of breast, colorectal, and lung cancers, we present results for local and distant stage lung and breast cancers since these represent the worst and best case scenarios for misclassification.

Lastly, we evaluated whether Medicare claims-based stage systematically misclassified stage for any patient group. We conducted multivariable logistic regression analyses to identify patient factors associated with stage misclassification, including age, race, gender, SEER registry, Charlson comorbidity score, and median census tract income. We also assessed whether patient factors were associated with the direction of stage misclassification (i.e., claims resulting in earlier or later stage classification vs. registry) in multivariable polytomous logistic regression analyses. Since the Charlson comorbidity score is measured in the year prior to cancer diagnosis, these sets of analyses were restricted to patients ages 66 and older.

Results

Table 1 presents socio-demographic and health characteristics of individuals diagnosed with breast, colorectal, and lung cancers. Among lung and breast cancer patients, there were similar distributions by age category. A greater proportion of colorectal cancer patients were ages 80 and older compared to other cancer sites. For all three cancer sites, the majority of the sample was White race. Among those ages 66 and older, comorbidity varied by cancer site, with 27.0% of lung cancer patients having a Charlson score of 2 or more compared to 19.5% and 13.8% of colorectal and breast cancer patients, respectively.

Aggregate distributions of stage

Figures 1A–1C depict the aggregate distribution of stage at diagnosis inferred from Medicare claims compared to SEER historic stage by cancer site. For all three cancers, Medicare claims-based stage classifications overestimated distributions of local and underestimated distributions of regional disease when compared to SEER data.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Discrepancies were greatest for lung cancer, including a substantially lower percentage of distant disease using the claims-based measure compared to SEER stage (22.9% vs. 47.4%, respectively).

Accuracy of Medicare claims for inferring stage

The sensitivity, specificity, PPV and NPV of the metastasis codes in Medicare claims relative to SEER stage is presented in Table 2. As shown in Table 2, none of the stage-specific accuracy measures simultaneously exceeded 80% for any of the cancer sites. Similarly, none of the "any metastases" measures simultaneously exceeded this threshold.

For patients with distant disease at diagnosis in the SEER data, the sensitivity, specificity and PPV of the claims were: lung (42.7%, 94.8%, 88.1%), breast (51.0%, 98.3%, 65.8%), and colorectal cancer (72.8%, 93.8%, 68.5%). In our sensitivity analyses using a less stringent definition of stage (i.e. only requiring a single metastasis code in physician claims), sensitivity was generally higher for regional and distant disease, but PPV was lower. Our sensitivity analyses using a 4-month window to classify stage in Medicare claims revealed similar findings (data not shown).

Comparison of survival based on SEER gold standard and Medicare claims-based stage at diagnosis

We also compared survival based on stage from Medicare claims to SEER stage for each cancer. Figures 2A and 2B illustrate a subset of these analyses for men with lung cancer and women with breast cancer. Survival based only on SEER stage is shown in each panel with a solid, dark blue line and was considered the gold standard. The shaded area around the lines represents the 95% confidence intervals. As shown in panel 1 of Figure 2A, the pink line represents survival for patients with agreement between local stage inferred from Medicare claims and SEER historic stage. Here, the pink line appears superimposed on SEER survival (dark blue line), demonstrating that when a patient's stage inferred from the Medicare claims corresponds to SEER stage, their survival is accurately represented. We also examined survival for patients with local stage derived from Medicare claims where the SEER data reported either regional stage (green line) or distant stage (brown line). For patients misclassified by Medicare claims as local stage, their observed survival was significantly poorer than for patients where the Medicare claims based measure corresponded to SEER stage ($p<0.001$) and the gold standard SEER survival.

Panel 2 of Figure 2A illustrates similar data for distant stage of lung cancer in men. Notably, among patients with disagreement for distant stage, survival is markedly better compared to SEER survival for distant stage (dark blue line) and these curves are represented by the green (SEER local) and brown (SEER regional) lines. For women with breast cancer, patterns of survival by Medicare claims-based stage and SEER data (Figure 2B, Panels 1 and 2) were similar to patterns for men with lung cancer. Misclassification of stage was significantly associated with –inaccurate estimates of stage-specific survival for all cancers ($p<0.001$).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Patient Factors Associated with Misclassification of Cancer Stage at Diagnosis Inferred from Medicare Claims

Table 3 presents results from multivariable logistic regressions evaluating the association between patient socio-demographic factors and stage disagreement between Medicare claims-based measures of stage at diagnosis and SEER data. In adjusted analysis, older age was consistently associated with greater stage misclassification in Medicare claims for all cancers ($p<0.01$ for all). Black patients were significantly more likely to have stage misclassification compared to other patients for lung ($p<0.001$) and breast cancers ($p<0.01$). Compared to those from the highest census-tract income areas, patients residing in lower census-tract income areas were more likely to have misclassification of stage by Medicare claims-based measures for all three cancers. Polytomous regression results indicated that misclassification for adults who were older, Black breast and lung cancer patients, and those living in lower census-tract income areas was in the direction of claims inferring an earlier stage of disease at diagnosis than the gold standard registry stage (Appendix Tables 2–4).

Discussion

In this study, we evaluated the validity of metastases codes in Medicare claims for inferring stage at diagnosis for breast, colorectal, and lung cancers. Overall performance of the metastases codes from claims data compared to the gold standard of SEER stage was poor and never simultaneously exceeded 80% for sensitivity, specificity, PPV and NPV for any stage for any cancer. Our findings are consistent with prior studies and demonstrate that use of claims alone to infer cancer stage at diagnosis will misclassify a significant number of patients and lead to a biased assessment of survival.[1,2,4] Use of claims alone to infer stage may also introduce bias in analyses where stage is evaluated as a confounder or effect modifier of other associations.

This study builds on prior research that has evaluated the validity of administrative claims to identify metastases in several important ways. Our assessment of stage-specific survival demonstrates that using Medicare claims to infer stage will inaccurately represent survival, with the greatest discrepancies for cancers with a higher proportion of advanced disease patients (e.g. lung and colorectal cancers). We also identified patient factors associated with systematic misclassification of inferred stage from Medicare claims. We found disproportionate stage misclassification in older cancer patients and individuals living in lower income census-tracts for all three cancers. Black lung and breast cancer patients were also more likely to have stage misclassification than their White counterparts. Additionally, our findings indicated that stage from claims was commonly misclassified as earlier stage of disease at diagnosis compared to the gold standard registry stage. Potential reasons for greater misclassification among these groups are likely tied to a combination of factors, such as health care setting, coding practice variations by type of institution, and individual-level discrepancies in quality of coding. Black patients and older adults are less likely to receive cancer treatment [18–21] and may also be less likely to receive complete diagnostic evaluation or staging. Possible explanations for this include limited access to care, transportation barriers, and geographic-based differences in availability of health care resources.[18–22]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Our analysis of the validity of Medicare claims for identifying metastasis focused on the time of diagnosis. We could not evaluate the validity of health claims for identifying metastasis, for inferring recurrence, or for calculating disease-free survival because there is not a gold standard measure of recurrence in population-based data. Although the period following diagnosis is when cancer patients are likely to be evaluated comprehensively, the poor performance we observed in the diagnosis period suggests that metastases codes alone in claims data will also have limited utility for accurately identifying metastasis for inferring cancer recurrence or disease progression. A recent population-based study using health maintenance organization (HMO) and the Cancer Care Outcomes Research and Surveillance Consortium (CanCORS) study data assessed disease-free status 14 and 60 months after diagnosis and concluded that no set of codes for metastasis were highly sensitive or specific among breast, lung, colorectal and prostate cancer patients.[23] Additionally, Warren, et al. used the SEER-Medicare data to examine indicators of recurrence after initial treatment and found that claims with metastasis codes and the dates of these claims were poor indicators of recurrence or its timing.[24] Therefore, our study provides additional evidence that relying on diagnosis codes for metastasis in claims will misclassify recurrence or disease-free survival among cancer patients.

While our findings suggest that there are substantive challenges to using administrative data to identify cancer metastasis and infer stage, they also offer insight into efforts that may improve data related to metastasis and cancer stage. Claims alone are unlikely to have sufficient detail to reliably identify metastatic disease, either at diagnosis or at disease progression. Administrative data-based algorithms can be modified to emphasize different aspects of validity (e.g., PPV).[25] Additionally, the development of electronic health records (EHR) and natural language processing tools offer opportunities to enhance information collected for cancer patients so that accurate stage and disease progression information could be reported in a consistent and comprehensive manner.[26,27]

Notably, this study had several strengths, including a large population-based sample, a cancer registry gold-standard definition of stage at diagnosis, the evaluation of validity for each stage as well as the broader category of “any metastases” and an assessment of the impact of stage misclassification on survival. We also conducted several sensitivity analyses to evaluate different Medicare claims algorithms for inferring stage, different windows of time following diagnosis, and different model specifications of misclassification. Despite these strengths, there were also limitations that should be noted. We were not able to provide information on groups that were excluded from the analyses, such as patients younger than age 65 or those receiving coverage through Medicare managed care. In addition, these findings may not be generalizable to patients treated outside of SEER regions.

In conclusion, our findings suggest that metastases codes should not be used to infer stage at diagnosis because this strategy results in substantial misclassification, particularly for certain socio-demographic groups. Furthermore, metastases codes alone from Medicare claims will likely have limited validity for inferring recurrence and estimating disease-free survival. Future research should utilize data sources with comprehensive stage information, such as registries or medical records, and make efforts to replicate this study’s findings among younger cancer patients and other cancer sites. Additionally, efforts to standardize the EHR

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

to collect cancer stage and recurrence information should be a priority for future health services research.

References

1. Cooper GS, Yuan Z, Stange KC, Amini SB, Dennis LK, Rimm AA. The utility of Medicare claims data for measuring cancer stage. Med Care. 1999 Jul; 37(7):706–11. [PubMed: 10424641]
2. Thomas SK, Brooks SE, Mullins CD, Baquet CR, Merchant S. Use of ICD-9 coding as a proxy for stage of disease in lung cancer. Pharmacoepidemiol Drug Saf. 2002 Dec; 11(8):709–13. [PubMed: 12512248]
3. Eichler AF, Lamont EB. Utility of administrative claims data for the study of brain metastases: a validation study. J Neurooncol. 2009 Dec; 95(3):427–31. Epub 2009 Jun 27. [PubMed: 19562256]
4. Nordstrom BL, Whyte JL, Stolar M, Mercaldi C, Kallich JD. Identification of metastatic cancer in claims data. Pharmacoepidemiology and drug safety. 2012; 21(S2):21–28. [PubMed: 22552976]
5. Gagnon B, Mayo NE, Laurin C, Hanley JA, McDonald N. Identification in administrative databases of women dying of breast cancer. J Clin Oncol. 2006 Feb 20; 24(6):856–62. [PubMed: 16484694]
6. Stokes ME, Thompson D, Montoya EL, Weinstein MC, Winer EP, Earle CC. Ten-year survival and cost following breast cancer recurrence: estimates from SEER-Medicare data. Value Health. 2008 Mar-Apr;11(2):213–20.10.1111/j.1524-4733.2007.00226.x [PubMed: 18380633]
7. Lamont EB, Herndon JE 2nd, Weeks JC, Henderson IC, Earle CC, Schilsky RL, Christakis NA. Cancer and Leukemia Group B. Measuring disease-free survival and cancer relapse using Medicare claims from CALGB breast cancer trial participants (companion to 9344). J Natl Cancer Inst. 2006 Sep 20; 98(18):1335–8. [PubMed: 16985253]
8. Warren JL, Klabunde CN, Schrag D, Bach PB, Riley GF. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. Med Care. 2002 Aug; 40(8 Suppl):IV-3–18.
9. Surveillance Epidemiology and End Results Program. [Accessed on April 30, 2014] Overview of the SEER Program. Available at: http://seer.cancer.gov/about/overview.html
10. Young, JL., Jr; Roffers, SD.; Ries, LAG.; Fritz, AG.; Hurlbut, AA., editors. SEER Summary Staging Manual - 2000: Codes and Coding Instructions. National Cancer Institute; Bethesda, MD: 2001. NIH Pub. No. 01-4969
11. Surveillance Epidemiology and End Results Program. [Accessed on April 30, 2014] SEER Summary Staging Manual – 2000. Available at: http://seer.cancer.gov/tools/ssm/
12. Surveillance Epidemiology and End Results Program. [Accessed on April 30, 2014] SEER Quality Improvement. Available at: http://seer.cancer.gov/qi/
13. The Applied Research Program. [Accessed on April 30, 2014] SEER-Medicare: How the SEER and Medicare Data are Linked. Available at: http://appliedresearch.cancer.gov/seermedicare/overview/linked.html
14. Warren JL, Klabunde CN, Schrag D, Bach PB, Riley GF. Overview of the SEER-Medicare Data: Content, Research Applications, and Generalizability to the United States Elderly Population. Med Care. 2002 Aug; 40(8 Suppl):3–18.
15. Lund JL, Stürmer T, Harlan LC, Sanoff HK, Sandler RS, Brookhart MA, Warren JL. Identifying Specific Chemotherapeutic Agents in Medicare Data: A Validation Study. Med Care. 2011 Nov 10. [Epub ahead of print].
16. Warren JL, Harlan LC, Fahey A, Virnig BA, Freeman JL, Klabunde CN, Cooper GS, Knopf KB. Utility of the SEER-Medicare data to identify chemotherapy use. Med Care. 2002 Aug; 40(8 Suppl):IV-55–61.
17. Klabunde CN, Legler JM, Warren JL, Baldwin LM, Schrag D. A refined comorbidity measurement algorithm for claims-based studies of breast, prostate, colorectal, and lung cancer patients. Ann Epidemiol. 2007 Aug; 17(8):584–90. Epub 2007 May 25. [PubMed: 17531502]
18. Vandergrift JL, Niland JC, Theriault RL, Edge SB, Wong YN, Loftus LS, Breslin TM, Hudis CA, Javid SH, Rugo HS, Silver SM, Lepisto EM, Weeks JC. Time to adjuvant chemotherapy for breast

NIH-PA Author Manuscript

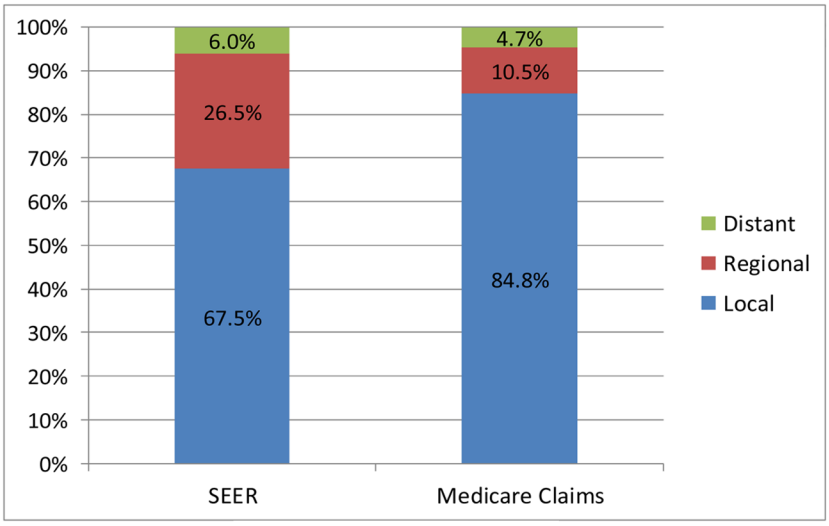
NIH-PA Author Manuscript

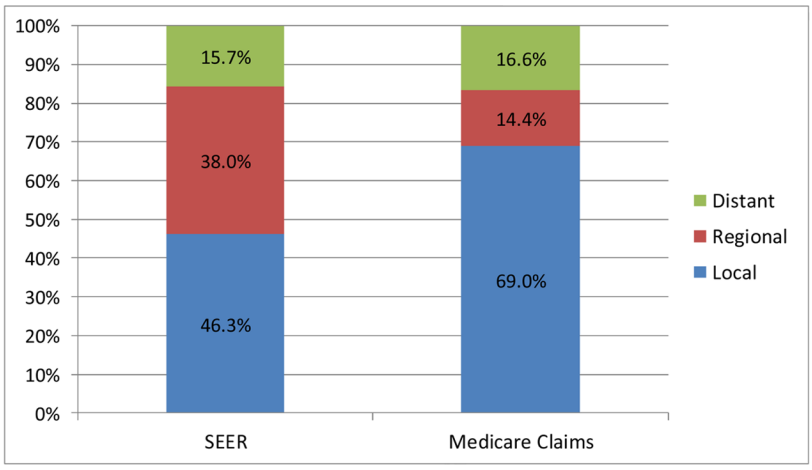
NIH-PA Author Manuscript

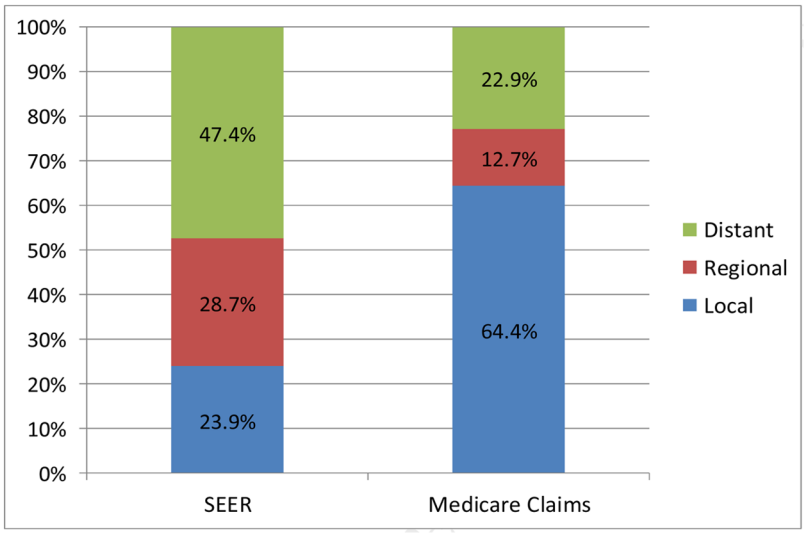
cancer in National Comprehensive Cancer Network institutions. J Natl Cancer Inst. 2013 Jan 16; 105(2):104–12. Epub 2012 Dec 21. 10.1093/jnci/djs506 [PubMed: 23264681]

19. Hines RB, Markossian TW. Differences in late-stage diagnosis, treatment, and colorectal cancer-related death between rural and urban African Americans and whites in Georgia. J Rural Health. 2012 Summer;28(3):296–305. Epub 2011 Aug 24. 10.1111/j.1748-0361.2011.00390.x [PubMed: 22757954]
20. Reeder-Hayes KE, Bainbridge J, Meyer AM, Amos KD, Weiner BJ, Godley PA, Carpenter WR. Race and age disparities in receipt of sentinel lymph node biopsy for early-stage breast cancer. Breast Cancer Res Treat. 2011 Aug; 128(3):863–71. Epub 2011 Feb 22. 10.1007/s10549-011-1398-1 [PubMed: 21340480]
21. Markossian TW, Hines RB. Disparities in late stage diagnosis, treatment, and breast cancer-related death by race, age, and rural residence among women in Georgia. Women Health. 2012; 52(4): 317–35.10.1080/03630242.2012.674091 [PubMed: 22591230]
22. Chien LC, Schootman M, Pruitt SL. The modifying effect of patient location on stage-specific survival following colorectal cancer using geosurvival models. Cancer Causes Control. 2013 Mar; 24(3):473–84. Epub 2013 Jan 10. 10.1007/s10552-012-0134-4 [PubMed: 23306551]
23. Hassett MJ, Ritzwoller DP, Taback N, Carroll N, Cronin AM, Ting GV, Schrag D, Warren JL, Hornbrook MC, Weeks JC. Validating Billing/Encounter Codes as Indicators of Lung, Colorectal, Breast, and Prostate Cancer Recurrence Using 2 Large Contemporary Cohorts. Med Care. 2012 Dec 6. [Epub ahead of print].
24. Warren JL, Mariotto A, Melbert D, Schrag D, Doria-Rose P, Penson D, Yabroff KR. Sensitivity of Medicare Claims to Identify Cancer Recurrence in Elderly Colorectal and Breast Cancer Patients. Med Care. 2013 Dec 26. [Epub ahead of print].
25. Chubak J, Yu O, Pocobelli G, Lamerato L, Webster J, Prout MN, Ulcickas Yood M, Barlow WE, Buist DS. Administrative data algorithms to identify second breast cancer events following early-stage invasive breast cancer. J Natl Cancer Inst. 2012 Jun 20; 104(12):931–40. Epub 2012 Apr 30. 10.1093/inci/dis233 [PubMed: 22547340]
26. Carrell DS, Halgrim S, Tran DT, Buist DS, Chubak J, Chapman WW, Savova G. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. Am J Epidemiol. 2014 Mar 15; 179(6):749–58. Epub 2014 Jan 30. 10.1093/aje/kwt441 [PubMed: 24488511]
27. Strauss JA, Chao CR, Kwan ML, Ahmed SA, Schottinger JE, Quinn VP. Identifying primary and recurrent cancers using a SAS-based natural language processing algorithm. J Am Med Inform Assoc. 2013 Mar-Apr;20(2):349–55. Epub 2012 Jul 21. 10.1136/amiaml-2012-000928 [PubMed: 22822041]

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript







NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

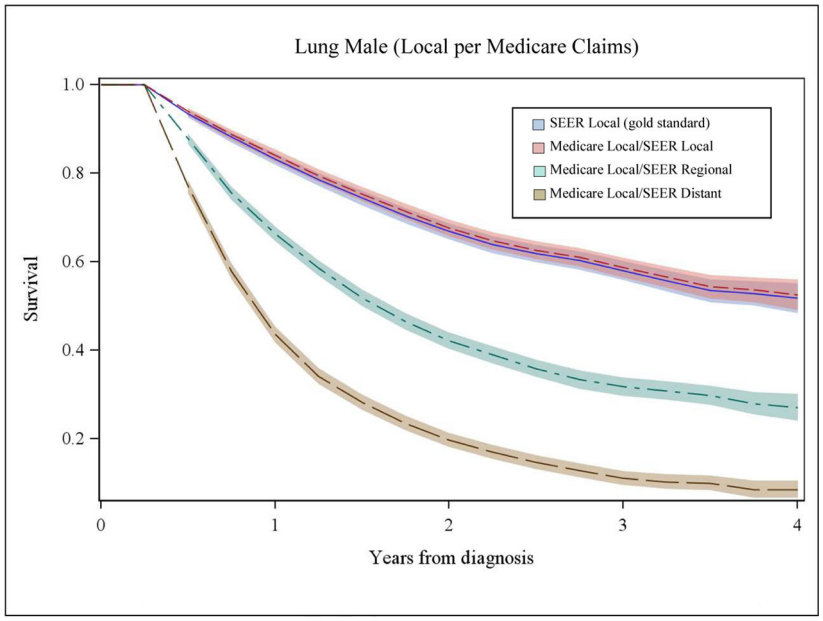
Figure 1.

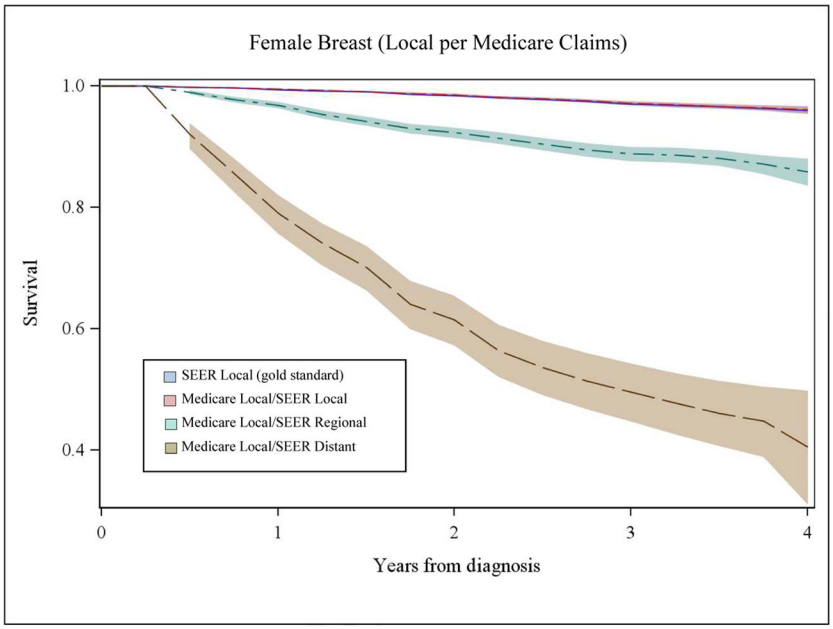
Figure 1A. Aggregate Distributions of Breast Cancer Stage at Diagnosis from SEER Data and Inferred from Medicare Claims

Figure 1B. Aggregate Distributions of Colorectal Cancer Stage at Diagnosis from SEER Data and Inferred from Medicare Claims

Figure 1C. Aggregate Distributions of Lung Cancer Stage at Diagnosis from SEER Data and Inferred from Medicare Claims

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

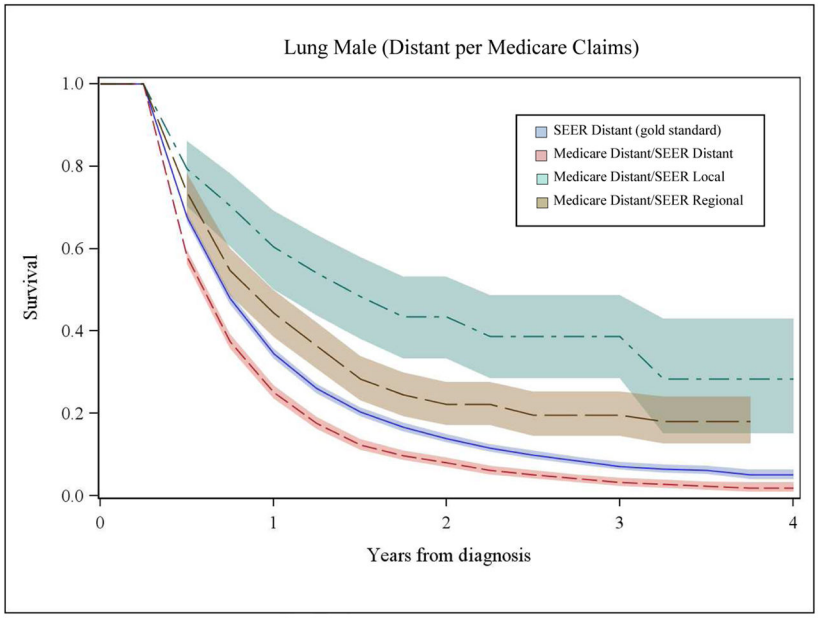




NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript



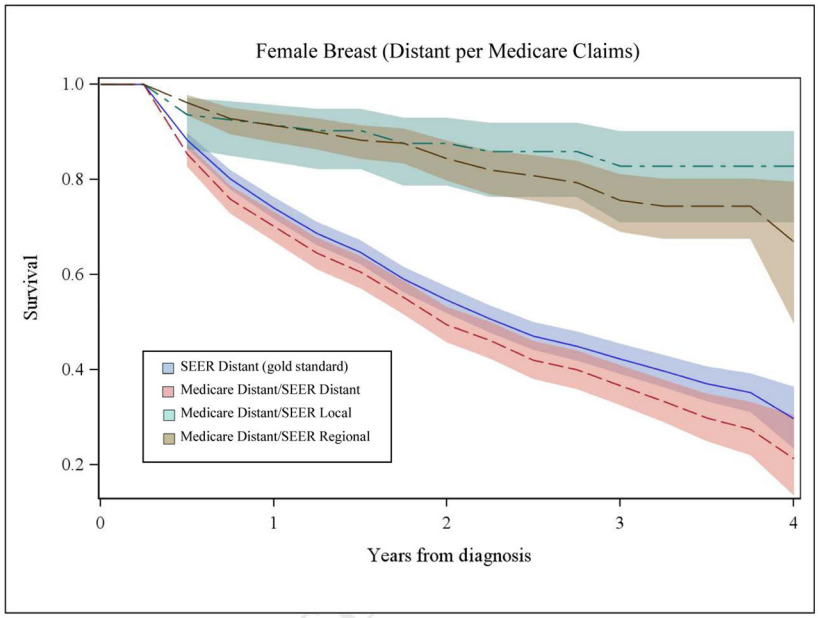


Figure 2.

Figure 2A, Panel 1. Comparison of Overall Survival for Men with Local Lung Cancer by Stage at Diagnosis Inferred from Medicare Claims to SEER Data

Figure 2B, Panel 1. Comparison of Overall Survival for Women with Local Breast Cancer by Stage at Diagnosis Inferred from Medicare Claims to SEER Data

Figure 2A, Panel 2. Comparison of Overall Survival for Men with Distant Lung Cancer by Stage at Diagnosis Inferred from Medicare Claims to SEER Data

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Figure 2B, Panel 2. Comparison of Overall Survival for Women with Distant Breast Cancer by Stage at Diagnosis Inferred from Medicare Claims to SEER Data

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Ann Epidemiol. Author manuscript; available in PMC 2015 September 01.

Table 1

Sociodemographic and Health Characteristics among Patients 65 and Older Diagnosed Between 2005–2007 with Breast, Colorectal, and Lung Cancers

	Breast		Colorectal		Lung	
	N	%	N	%	N	%
Age						
65–69	7,098	26.2	5,178	21.4	7,418	25.9
70–74	6,190	22.8	5,136	21.2	7,429	25.9
75–79	5,772	21.3	5,296	21.9	6,895	24.0
80+	8,083	29.8	8,606	35.5	6,951	24.2
Race						
White	23,876	88.0	20,655	85.3	24,838	86.6
Black	1,878	6.9	1,987	8.2	2,330	8.1
Other/Unknown	1,389	5.1	1,574	6.5	1,525	5.3
Gender						
Male	0	0	10,928	45.1	14,315	49.9
Female	27,143	100.0	13,288	54.9	14,378	50.1
Charlson comorbidity score in the year prior to diagnosis *						
0	15,941	63.5	12,469	55.2	11,202	42.0
1	5,699	22.7	5,703	25.3	8,252	31.0
2+	3,453	13.8	4,404	19.5	7,206	27.0
Total	25,093	100.0	22,576	100.0	26,660	100.0
2000 census tract median income quartile						
<$34,456	5,989	22.1	6,165	25.5	7,823	27.3
$34,456 – $45,760	6,658	24.5	6,105	25.2	7,222	25.2
$45,761 – $61,234	6,987	25.7	5,961	24.6	7,018	24.5
$61,235+	7,470	27.5	5,944	24.5	6,576	22.9
Missing	39	0.1	41	0.2	54	0.2
Total	27,143	100.0	24,216	100.0	28,693	100.0

* Charlson comorbidity score was calculated among adults 66 and older and cancer was not considered a comorbidity in this study.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 2

Accuracy of Stage Inferred from Medicare Claims to Identify Stage Reported in the SEER Data (Gold Standard)

	SEER Historic Stage		Stage Inferred from Metastasis Codes in Medicare Claims							
		N	%							
Cancer site			Local	Regional	Distant	Any metastases*	Sensitivity	Specificity	PPV	NPV
Breast	Local	18,310	98.6	0.9	0.5	1.4	98.6	43.9	78.5	93.8
	Regional	7,191	60.0	35.3	4.7	40.0	35.3	98.4	88.8	80.8
	Distant	1,642	39.3	9.7	51.0	60.7	51.0	98.3	65.8	96.9
	Any metastases*	8,833	56.1	30.5	13.3	43.9	43.9	98.6	93.8	78.5
Colorectal	Local	11,213	96.8	1.0	2.2	3.2	96.8	55.0	65.0	95.2
	Regional	9,213	55.0	34.0	11.0	45.0	34.0	97.6	89.7	70.6
	Distant	3,790	20.6	6.6	72.8	79.4	72.8	93.8	68.5	94.9
	Any metastases*	13,003	45.0	26.0	29.0	55.0	55.0	96.8	95.2	65.0
Lung	Local	6,850	94.2	3.2	2.6	5.8	94.3	45.0	34.9	96.1
	Regional	8,249	75.0	17.6	7.4	25.0	17.6	89.3	40.0	72.9
	Distant	13,594	42.9	14.4	42.7	57.1	42.7	94.8	88.1	64.7
	Any metastases*	21,843	55.0	15.6	29.3	45.0	45.0	94.3	96.1	34.9

Note: PPV = positive predictive value, NPV = negative predictive value;

*Any metastases combines regional and distant stage.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 3

Multivariable Analyses of Patient Factors Associated with Misclassification of Cancer Stage at Diagnosis Inferred from Medicare Claims

	Breast (N =25,093)				Colorectal (N=22,576)				Lung (N=26,660)			
		Adjusted OR ∞		Wald p-value		Adjusted OR ∞		Wald p-value		Adjusted OR ∞		Wald p-value
	% Disagree	OR	95% CI		% Disagree	OR	95% CI		% Disagree	OR	95% CI	
Age group at diagnosis												
66–69 (ref)	20.0	1.00		p=0.0033	29.6	1.00		p=0.0054	50.4	1.00		p<0.0001
70–74	21.0	1.07	0.98 – 1.17		29.7	1.01	0.92 – 1.10		51.0	1.04	0.97 – 1.11	
75–79	20.1	1.01	0.92 – 1.11		31.0	1.07	0.98 – 1.18		51.7	1.08	1.00 – 1.16	
80+	22.2	1.15	1.06 – 1.26		32.1	1.13	1.04 – 1.23		56.1	1.29	1.21 – 1.39	
Race				p=0.0015				p=0.7642				p<0.0001
White (ref)	20.6	1.00			30.9	1.00			51.7	1.00		
Black	27.6	1.24	1.10 – 1.40		31.1	0.96	0.86 – 1.08		58.3	1.31	1.19 – 1.44	
Other/Unknown	19.4	1.08	0.92 – 1.26		30.0	0.98	0.86 – 1.11		54.4	1.05	0.93 – 1.19	
Sex								p=0.8429				p<0.0001
Male (ref)					30.6	1.00			54.0	1.00		
Female	21.0				31.1	1.01	0.95 – 1.07		50.7	0.87	0.83 – 0.91	
Charlson comorbidity score in the year prior to diagnosis				p=0.1393				p=0.1032				p=0.2412
0 (ref)	20.4	1.00			31.3	1.00			52.7	1.00		
1	21.5	1.04	0.96 – 1.12		30.3	0.94	0.88 – 1.01		51.6	0.95	0.90 – 1.01	
2+	23.0	1.09	1.00 – 1.20		30.3	0.93	0.87 – 1.01		52.8	0.98	0.93 – 1.04	
2000 census tract median income quartile				p<0.0001				p=0.0010				p=0.0028
<$34,456	24.9	1.30	1.18 – 1.44		32.3	1.19	1.09 – 1.31		53.9	1.15	1.06 – 1.24	
$34,456 – $45,760	20.4	1.07	0.98 – 1.17		30.7	1.09	1.00 – 1.19		53.4	1.12	1.05 – 1.21	
$45,761 – $61,234	20.0	1.04	0.95 – 1.13		31.4	1.12	1.03 – 1.22		51.8	1.07	1.00 – 1.15	
$61,235+ (ref)	19.3	1.00			29.2	1.00			50.1	1.00		

Note: OR= odds ratio, CI=Confidence Interval; ∞Model also adjusted for location of SEER registry in addition to variables shown; ref means reference group; Sample restricted to 66 and older who had part A/B non-HMO coverage to allow for evaluation of non-cancer comorbidity in year prior to diagnosis; Some observations were not used in models due to missing values.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Appendix Table 1

Diagnosis Codes used to Identify Metastases in Medicare Claims and Infer Stage

Site	Regional Metatases	Distant Metastases
Lung	196.1 Intrathoracic lymph nodes- Bronchopulmonary Intercostal Mediastinal Tracheobronchial 197.0 Lung Bronchus 197.1 Mediastinum 197.2 Pleura 197.3 Other respiratory organs-Trachea	Lymph nodes 196.0, 196.2–196.9 Abdomen 197.4–197.6, 197.8 Liver 197.7 Bone 198.5 Brain 198.3 Carcinomatosis 199.0 Other sites 198.0–198.2, 198.4,198.6–198.8x
Colorectal	196.2 Intra-abdominal lymph nodes Intestinal Mesenteric Retroperitoneal 197.5 Large intestine and rectum	Lymph nodes 196.0, 196.1, 196.3–196.9 Lung 197.0–197.3 Abdomen 197.4, 197.6, 197.8 Liver 197.7 Bone 198.5 Brain 198.3 Carcinomatosis 199.0 Other sites 198.0–198.2, 198.4, 198.6–198.8x
Breast	196.3 Lymph nodes of axilla and upper limb Brachial Epitrochlear Infraclavicular Pectoral 198.81 Breast	Lymph nodes 196.1, 196.2, 196.4–196.9 Lung 197.0–197.3 Abdomen 197.4–197.6, 197.8 Liver 197.7 Bone 198.5 Brain 198.3 Carcinomatosis 199.0 Other sites 198.0–198.2, 198.4, 198.6–198.7, 198.82, 198.89

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Appendix Table 2

Patient Factors Associated with Misclassification of Breast Cancer Stage at Diagnosis Inferred from Medicare Claims

	% Disagree	Claims classifies as earlier stage vs. Registry		Claims classifies as later stage vs. Registry	
		Adjusted OR	95% CI	Adjusted OR	95% CI
Age*					
66–69 (ref)	20.0	1.00		1.00	
70–74	21.0	1.07	(0.97–1.18)	1.06	(0.82–1.37)
75–79	20.1	1.02	(0.92–1.12)	0.97	(0.75–1.26)
80+	22.2	1.17	(1.07–1.28)	1.04	(0.82–1.32)
Race***					
White (ref)	20.6	1.00		1.00	
Black	27.6	1.30	(1.14–1.47)	0.82	(0.57–1.18)
Other/Unknown	19.4	1.06	(0.90–1.25)	1.18	(0.77–1.81)
Gender					
Male (ref)		---	---	---	---
Female	21.0	---	---	---	---
Charlson comorbidity score in the year prior to diagnosis					
0 (ref)	20.4	1.00		1.00	
1	21.5	1.02	(0.95–1.11)	1.14	(0.93–1.40)
2+	23.0	1.07	(0.98–1.18)	1.28	(1.00–1.62)
2000 census tract median income quartile***					
<$34,456	24.9	1.30	(1.18–1.44)	1.31	(1.01–1.70)
$34,456 – $45,760	20.4	1.09	(0.99–1.20)	0.96	(0.75–1.24)
$45,761 – $61,234	20.0	1.06	(0.97–1.16)	0.88	(0.69–1.12)
$61,235+ (ref)	19.3	1.00		1.00	

Note: Males were excluded from analyses of breast cancer patients; OR=odds ratio, CI=Confidence Interval; ∞Model also adjusted for location of SEER registry in addition to variables shown; ref means reference group; Sample restricted to 66 and older who had part A/B non-HMO coverage to allow for evaluation of non-cancer comorbidity in year prior to diagnosis; Some observations were not used in models due to missing values; Wald p-value

*p<0.05,

** p<0.01, *** p<0.001.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Appendix Table 3

Patient Factors Associated with Misclassification of Colorectal Cancer Stage at Diagnosis Inferred from Medicare Claims

	% Disagree	Claims classifies as earlier stage vs. Registry		Claims classifies as later stage vs. Registry	
		Adjusted OR	95% CI	Adjusted OR	95% CI
Age*					
66–69 (ref)	29.6	1.00		1.00	
70–74	29.7	1.01	(0.91–1.11)	1.00	(0.83–1.21)
75–79	31.0	1.06	(0.97–1.17)	1.11	(0.93–1.34)
80+	32.1	1.13	(1.03–1.24)	1.14	(0.96–1.35)
Race					
White (ref)	30.9	1.00		1.00	
Black	31.1	0.96	(0.85–1.08)	0.98	(0.78–1.23)
Other/Unknown	30.0	0.94	(0.82–1.08)	1.14	(0.89–1.47)
Gender**					
Male (ref)	30.6	1.00		1.00	
Female	31.1	0.97	(0.91–1.03)	1.19	(1.06–1.34)
Charlson comorbidity score in the year prior to diagnosis					
0 (ref)	31.3	1.00		1.00	
1	30.3	0.92	(0.86–0.99)	1.05	(0.92–1.20)
2+	30.3	0.93	(0.86–1.01)	0.94	(0.81–1.10)
2000 census tract median income quartile**					
<$34,456	32.3	1.22	(1.11–1.35)	1.08	(0.90–1.29)
$34,456 – $45,760	30.7	1.12	(1.02–1.22)	0.99	(0.84–1.18)
$45,761 – $61,234	31.4	1.15	(1.05–1.26)	1.01	(0.86–1.19)
$61,235+ (ref)	29.2	1.00		1.00	

Note: OR= odds ratio, CI=Confidence Interval; ∞Model also adjusted for location of SEER registry in addition to variables shown; ref means reference group; Sample restricted to 66 and older who had part A/B non-HMO coverage to allow for evaluation of non-cancer comorbidity in year prior to diagnosis; Some observations were not used in models due to missing values; Wald p-value

*p<0.05,

**p<0.01,

*** p<0.001.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Ann Epidemiol. Author manuscript; available in PMC 2015 September 01.

Appendix Table 4

Patient Factors Associated with Misclassification of Lung Cancer Stage at Diagnosis Inferred from Medicare Claims

	% Disagree	Claims classifies as earlier stage vs. Registry		Claims classifies as later stage vs. Registry	
		Adjusted OR	95% CI	Adjusted OR	95% CI
Age***					
66–69 (ref)	50.4	1.00		1.00	
70–74	51.0	1.04	(0.97–1.12)	0.96	(0.80–1.16)
75–79	51.7	1.09	(1.01–1.17)	0.98	(0.81–1.18)
80+	56.1	1.32	(1.23–1.42)	0.93	(0.77–1.14)
Race***					
White (ref)	51.7	1.00		1.00	
Black	58.3	1.34	(1.21–1.47)	1.00	(0.77–1.30)
Other/Unknown	54.4	1.04	(0.92–1.18)	1.19	(0.86–1.66)
Gender***					
Male (ref)	54.0	1.00		1.00	
Female	50.7	0.87	(0.82–0.91)	0.95	(0.83–1.09)
Charlson comorbidity score in the year prior to diagnosis					
0 (ref)	52.7	1.00		1.00	
1	51.6	0.95	(0.89–1.00)	1.02	(0.87–1.20)
2+	52.8	0.98	(0.92–1.04)	1.06	(0.90–1.25)
2000 census tract median income quartile**					
<$34,456	53.9	1.15	(1.06–1.24)	1.08	(0.88–1.33)
$34,456 – $45,760	53.4	1.15	(1.06–1.23)	0.87	(0.71–1.07)
$45,761 – $61,234	51.8	1.08	(1.00–1.16)	1.01	(0.84–1.23)
$61,235+ (ref)	50.1	1.00		1.00	

Note: OR= odds ratio, CI=Confidence Interval; ∞Model also adjusted for location of SEER registry in addition to variables shown; ref means reference group; Sample restricted to 66 and older who had part A/B non-HMO coverage to allow for evaluation of non-cancer comorbidity in year prior to diagnosis; Some observations were not used in models due to missing values; Wald p-value

*p<0.05,

**p<0.01,

*** p<0.001.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript