



Published in final edited form as:

Ann Epidemiol. 2015 April ; 25(4): 297–300. doi:10.1016/j.annepidem.2015.01.005.

Development of a claims-based algorithm to identify colorectal cancer recurrence

Anjali D. Deshpande¹, Mario Schootman², and Allese Mayer²

¹Division of General Medical Sciences, School of Medicine, Washington University in St. Louis, St. Louis, MO

²Department of Epidemiology, College for Public Health and Social Justice, Saint Louis University, St. Louis, MO

Abstract

Purpose—To examine the validity of claims data to identify CRC recurrence and determine the extent to which misclassification of recurrence status affects estimates of its association with overall survival in a population-based administrative database.

Methods—We calculated the accuracy of claims data relative to medical records from one large tertiary hospital to identify CRC recurrence. We estimated the effect of misclassifying recurrence on survival by applying these findings to the linked SEER-Medicare data.

Results—Of 174 eligible CRC patients identified through medical records, 32 (18.4%) had a recurrence. A claims-based algorithm of secondary malignancy codes yielded a sensitivity of 81% and specificity of 99% for identifying recurrence. Agreement between data sources was almost perfect (kappa: 0.86). In a model unadjusted for misclassification, CRC patients with recurrence were 3.04 times (95% CI: 2.92 – 3.17) more likely to die of any cause than those without recurrence. In the corrected model, CRC patients with recurrence were 3.47 times (95% CI 3.06 - 4.14) more likely to die than those without recurrence.

Conclusion—Identifying recurrence in CRC patients using claims data is feasible with moderate sensitivity and high specificity. Future studies can use this algorithm with SEER-Medicare data to study treatment patterns and outcomes of CRC patients with recurrence.

Keywords

claims-based algorithm; colorectal cancer; recurrence; misclassification

© 2015 Elsevier Inc. All rights reserved.

Corresponding author: Anjali Deshpande, PhD, MPH, Division of General Medical Sciences, Washington University School of Medicine, 600 South Taylor Avenue, Campus Box 8005, St. Louis, MO 63110, adeshpan@dom.wustl.edu, Phone: 314-286-0148, Fax: 314-286-1919.

The authors do not have any potential conflicts of interest to disclose.

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.

Background

Colorectal cancer (CRC) is the third most common cancer in the US. About 75% of CRC cases can be treated with curative resection, however approximately 50% of these patients will develop recurrent disease, most within 2 years.[1] Many CRC patients will die of their recurrent disease unless detected early enough to receive curative treatment.[2-4]

Studies have identified recurrence through self-report, medical record review, and claims data. Administrative claims data are ideally suited to conduct large population-based studies, but are hampered by lack of information about their ability to accurately identify recurrence. Being able to accurately identify recurrences allows researchers to study the “experiences and outcomes of patients with recurrent cancer, better control for the impact of recurrent disease on survival, and realize the full potential of administrative databases for comparative effectiveness research.”[5]

Previous studies to develop recurrence algorithms using administrative data observed low sensitivities which could lead to a high degree of misclassification and biased estimates of exposure-disease relationships.[5-12] As a result, these algorithms are of limited value. Our purpose was to develop an acceptable claims-based algorithm to identify recurrence in CRC patients and to determine the algorithm's utility in studying recurrence in a large population-based administrative database.

Methods

This study has two components: 1) accuracy of claims data relative to medical records to identify recurrence following CRC, and 2) estimation of the effect of misclassifying recurrence on overall survival in the linked Surveillance, Epidemiology, and End Results (SEER)-Medicare data. This study was approved by Washington University's Institutional Review Board.

1: Accuracy of claims data

Data Sources and Abstraction—We used two data sources: 1) clinical and tumor data from Barnes-Jewish Hospital (BJH) Oncology Data Services (ODS) that are routinely obtained from medical records for reporting to the statewide cancer registry and 2) all inpatient and outpatient hospital billing data from BJH's finance office for each CRC patient from the date of admission for their curative resection until the end of the follow-up period, December 31, 2010. Sociodemographic and clinical characteristics were obtained from ODS.

Study Population—To increase applicability, we included patients with the same characteristics in both parts of the study. We included patients aged 65 years and older who were diagnosed with a first primary CRC (sequence number 00, ICD-9-CM codes: 153.0-154.1) between January 1, 2005 and December 31, 2009, who were not diagnosed with an hereditary or familial cancer syndrome, and had curative resection of their primary tumor within 4 months of diagnosis from ODS (N=381). We excluded CRC patients with in-situ or stage IV disease (n=11); without curative resection at BJH (n=38); who were not

Medicare Part A and B fee-for-service enrollees or who were enrolled in managed care (n=61); who had a secondary malignant neoplasm diagnosis within 3 months of curative surgery (n=1); and persons who did not receive continuous follow-up oncology care and medical surveillance for at least 12 months post-surgery at BJH (n=96). The final study sample included 174 patients. We obtained billing data, including all diagnosis, treatment, and procedure codes, for these patients up to December 31, 2010.

Recurrence Algorithms—We defined recurrence as the development of new local recurrent or distant metastatic lesions after initial curative surgery.[13] The ODS data identified recurrence from medical records using physician notes, laboratory, pathology, imaging reports, or letter by an external physician indicating recurrence. We identified recurrence from claims data using three separate algorithms: (1) the presence of any diagnosis code indicating a secondary malignant neoplasm three or more months after the index surgery, including the ICD-9-CM diagnosis codes 196.2, 197, 197.0-197.8, 198.0-198.8, 198.82, and 198.89 [8]; (2) the presence of any treatment or procedure codes that indicated restarting or new chemotherapy, radiation or surgical treatments [7, 14]; and (3) algorithm 1 and 2. Earle and colleagues[7] suggest that modern treatment regimens for CRC are completed within 6 to 8 months of surgery; that most relapses occur within the first 24 months after diagnosis; and that “relapse may be indicated in a patient who received chemotherapy 16 months or more after initial treatment and/or radiation therapy 12 months or more after initial treatment.” We therefore looked at 2 possible treatment algorithms of recurrence as: 1) chemotherapy and/or radiation that started 8 months or more after surgery; and 2) chemotherapy 16 months or more after surgery and/or radiation treatment 12 months or more after surgery. The codes used were based on ICD9 and HCPCS codes as described by Warren.[15]

Statistical Analysis—The ODS data, enhanced by medical record abstraction, was considered the gold standard against which claims data were compared. We excluded ICD-9-CM diagnosis code 197.5 (secondary malignant neoplasm of the large intestine and rectum) from the algorithm because we felt this code may inadequately distinguish between a recurrence and the existing primary CRC as others have previously done in developing recurrence algorithms.[5, 8] ICD-9-CM diagnosis code 196.2, secondary and unspecified malignant neoplasm of intra-abdominal lymph nodes, was excluded due to inconsistent coding.

We calculated sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and associated 95% confidence intervals (CI) for each of the three algorithms versus ODS data. Sensitivity is the proportion of patients having a recurrence identified through claims data among those with a recurrence based on the ODS. Specificity is the proportion of patients free of recurrence until death or last follow-up from claims data among those without recurrence based on ODS. PPV was the proportion of patients identified by claims data with CRC recurrence who had a recurrence based on ODS. NPV was defined as the proportion of patients identified by claims data without recurrence who did not have a recurrence based on ODS. Agreement between the two data sources was assessed using Cohen's kappa with the commonly used adjectival ratings to interpret the

results: 0.80 to 1.00 (almost perfect agreement), 0.60 to 0.79 (substantial agreement), 0.40 to 0.59 (moderate agreement), 0.20 to 0.39 (fair agreement), and 0.00 to 0.19 (poor agreement). [16] Because kappa is affected by prevalence (i.e., recurrence), we also calculated the prevalence-bias adjusted kappa (PABAK).

2. Misclassified recurrence and survival

We obtained data from an existing linkage of 2000-2005 SEER program data of 12 registries with 1999-2005 Medicare claim files from the Centers for Medicare and Medicaid Services. We again included patients aged 65 years and older who were diagnosed with a first primary CRC. We used the aforementioned criteria to exclude CRC patients.

Statistical analysis—We used proportional hazard models to determine the unadjusted and adjusted hazard ratios of recurrence (regardless of time since surgery) on overall survival. We only report the results from the proportional hazard models because competing-risk models only marginally changed the hazard ratios. We also quantified the effects of misclassifying a dichotomous variable (i.e., recurrence)[17] by reconstructing the data that would have been observed had recurrence been correctly classified, given its sensitivity and specificity. Because the true sensitivity and specificity are seldom known, two trapezoid probability distributions are specified, using the aforementioned sensitivities we observed. We used 20,000 repetitions to randomly sample sensitivity and specificity from these distributions to obtain 20,000 estimates of the back-calculated hazard ratios, including the 2.5 percentile, the median, and the 97.5 percentile. For additional details about these calculations, see Lash and Fox.[18] Sensitivity and specificity of recurrence were assumed to be misclassified independently from a patient's vital status. We used the `episens` command in Stata (version 12.1) to adjust the observed hazard ratio for misclassification bias.[19]

Results

174 CRC patients were identified from ODS data, 32 (18.4%) of whom had recurrence based on medical record abstraction (Table 1).

1. Accuracy of using claims data

Table 2 shows moderately high sensitivity (81.3%) and very high specificity (99.3%), PVP (96.3%), and NPV (95.9%) for the ICD9 secondary malignancy code-based algorithm. Algorithms using treatment or procedure codes alone showed very low sensitivity. The algorithm combining ICD9 secondary malignancy codes with treatment or procedure codes did not identify any additional recurrences compared to the ICD9 secondary malignancy code algorithm alone (data not shown). Percentage of agreement was 96.0 percent. Kappa and PABAK indicated almost perfect agreement, 0.86, and 0.92, respectively.

2. Misclassified recurrence and survival

Patients with a recurrence based on our claims-based algorithm were 3.04 times (95% CI: 2.92-3.17) more likely to die than those without recurrence. Guided by the 95% CI of our observed sensitivity and specificity, we used values for sensitivity (minimum: 0.60, mode 1:

0.75, mode 2: 0.85, maximum: 0.93) and specificity (minimum: 0.95, mode 1: 0.98, mode 2: 0.99, maximum: 1.00) to describe the trapezoid probability distributions. Adjusted for misclassification, patients with a recurrence were 3.47 times (2.5%: 3.06, 97.5%: 4.14) more likely to die than those without a recurrence.

Because our algorithm might have different sensitivities in different patient populations or clinical settings, we examined its effect on the hazard ratio by varying the sensitivity of the algorithm, keeping the specificity constant. Lower sensitivities of the algorithm underestimated the true hazard ratio more (Table 3).

Discussion

The potential to investigate subsequent CRC events in administrative claims data is enormous. Several studies developed algorithms to identify recurrence using combinations of inpatient secondary malignancy codes and/or treatment codes for various cancers.[5-12] We developed an algorithm of ICD-9 secondary malignancy codes to identify recurrence following CRC surgery with acceptable sensitivity and very high specificity. Surprisingly, the best algorithm was comprised of secondary malignancy codes only; additional treatment or procedure codes did not improve its accuracy.

Several other studies have examined the use of claims-based data to identify recurrence after a primary colorectal cancer. [5, 6, 12, 20] The work by Anaya and colleagues[12] specifically looked at using claims data to identify CRC metastases to the liver and McClish et. al.[6] combined recurrences with second primaries and had only a small number of recurrences (n=15) thereby making these findings difficult to interpret. Warren and colleagues[20] examined the sensitivity of Medicare claims data for treatments to identify recurrence in elderly CRC patients (stage II/III) who later died from their cancer. They concluded that relying on treatment claims in Medicare data as an indicator of recurrence would result in significant underestimation of recurrences, especially among older patients and female CRC cases. Our observed low sensitivities of treatment algorithms for identifying CRC recurrences provides further support for their conclusion that treatment claims may lead to underestimation of CRC recurrences. Our study was most comparable to the study by Hassett.[5] They evaluated the validity of secondary malignancy codes and chemotherapy codes to identify recurrence after definitive treatment for stage I-III colorectal cancer in two cohorts of patients, one from CanCORS and the other from an HMO-based Cancer Research Network. The sensitivities ranged from 56% to 74% and, contrary to our findings, they found that sensitivities improved to a range of 75-83% when combining secondary malignancy codes with chemotherapy codes. Differences in the included secondary malignancy ICD9 codes between studies may explain slightly higher sensitivity for our ICD9 code-based algorithm. Additionally, their chemotherapy claims algorithm differs from our treatment algorithm in that the timelines used were different, they only included chemotherapy while we looked at both chemotherapy and radiation therapy, and their algorithm of chemotherapy codes included a greater variety of codes including National Drug Codes which were not available in our data. These differences in study population and methodology may account for observed differences in our results.

Our estimated effect of misclassifying recurrence on survival gives researchers an important and easy to use secondary malignancy ICD9 code-based tool for studying recurrence and mortality outcomes in the SEER-Medicare data. Addressing misclassification is important when using ICD-9 code-based algorithms to identify cancer recurrence in administrative databases.

Study limitations are 1) potential inaccuracy of medical records as the gold standard; 2) excluding a large portion of patients at the study institution because they did not receive at least 12 months of follow up at the study institution; and 3) a relatively small sample size. Realistically, validation studies that compare administrative data with medical records data can only feasibly be done in small settings.

In conclusion, identifying CRC patients with recurrence using administrative data is feasible. Future studies can use the SEER-Medicare data and the proposed methodology for adjusting for misclassification to study the epidemiology, treatment patterns, and outcomes of CRC patients with recurrence.

Acknowledgments

We thank the Alvin J. Siteman Cancer Center at Barnes-Jewish Hospital and Washington University School of Medicine in St. Louis, Missouri, for the use of the Health Behavior, Communication, and Outreach Core. This study used the linked SEER-Medicare database. The interpretation and reporting of these data are the sole responsibility of the authors. The authors acknowledge the efforts of the Applied Research Program, NCI; the Office of Research, Development and Information, CMS; Information Management Services (IMS), Inc.; and the Surveillance, Epidemiology, and End Results (SEER) Program tumor registries in the creation of the SEER-Medicare database. We also thank Ms. Sida Yan for her assistance in medical record abstraction and Mr. Jim Struthers for his assistance in data management.

Role of the Funding Sources: This research was supported in part by grants from the National Cancer Institute (CA91842, CA137750). The funders did not have any role in the design of the study; the analysis or interpretation of the data; the decision to submit the manuscript for publication; or the writing of the manuscript.

Literature Cited

1. Figueredo A, Rumble RB, Maroun J, Earle CC. Follow-up of patients with curatively resected colorectal cancer: a practice guideline. *BMC Cancer*. 2003; 3:26. [PubMed: 14529575]
2. Sargent DJ, Wieand HS, Haller DG, Gray R, Benedetti JK, Buyse M, Labianca R, Seitz JF, O'Callaghan CJ, Francini G, et al. Disease-free survival versus overall survival as a primary end point for adjuvant colon cancer studies: individual patient data from 20,898 patients on 18 randomized trials. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. 2005; 23(34):8664–8670. [PubMed: 16260700]
3. Kievit J. Follow-up of patients with colorectal cancer: numbers needed to test and treat. *Eur J Cancer*. 2002; 38(7):986–999. [PubMed: 11978524]
4. Pfannschmidt J, Dienemann H, Hoffmann H. Surgical resection of pulmonary metastases from colorectal cancer: a systematic review of published series. *The Annals of thoracic surgery*. 2007; 84(1):324–338. [PubMed: 17588454]
5. 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
6. McClish D, Penberthy L, Pugh A. Using Medicare claims to identify second primary cancers and recurrences in order to supplement a cancer registry. *J Clin Epidemiol*. 2003; 56(8):760–767. [PubMed: 12954468]

7. Earle CC, Nattinger AB, Potosky AL, Lang K, Mallick R, Berger M, Warren JL. Identifying cancer relapse using SEER-Medicare data. *Med Care*. 2002; 40(8 Suppl):IV-75–81.
8. Lamont EB, Herndon JE 2nd, Weeks JC, Henderson IC, Earle CC, Schilsky RL, Christakis NA. Cancer, 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; 98(18):1335–1338. [PubMed: 16985253]
9. Lamont EB, Herndon JE 2nd, Weeks JC, Henderson IC, Lilenbaum R, Schilsky RL, Christakis NA. Criterion validity of Medicare chemotherapy claims in Cancer and Leukemia Group B breast and lung cancer trial participants. *J Natl Cancer Inst*. 2005; 97(14):1080–1083. [PubMed: 16030306]
10. Eichler AF, Lamont EB. Utility of administrative claims data for the study of brain metastases: a validation study. *Journal of neuro-oncology*. 2009; 95(3):427–431. [PubMed: 19562256]
11. 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; 104(12):931–940. [PubMed: 22547340]
12. Anaya DA, Becker NS, Richardson P, Abraham NS. Use of administrative data to identify colorectal liver metastasis. *The Journal of surgical research*. 2012; 176(1):141–146. [PubMed: 21962740]
13. Tsai HL, Chu KS, Huang YH, Su YC, Wu JY, Kuo CH, Chen CW, Wang JY. Predictive factors of early relapse in UICC stage I-III colorectal cancer patients after curative resection. *Journal of surgical oncology*. 2009; 100(8):736–743. [PubMed: 19757443]
14. Warren JL, Yabroff KR, Meekins A, Topor M, Lamont EB, Brown ML. Evaluation of trends in the cost of initial cancer treatment. *J Natl Cancer Inst*. 2008; 100(12):888–897. [PubMed: 18544740]
15. Warren JL, Yabroff KR, Meekins A, Topor M, Lamont EB, Brown M. Evaluation of trends in the cost of initial cancer treatment. *J Natl Cancer Inst*. 2008; 100(12):888–897. [PubMed: 18544740]
16. Landis J, Koch G. The measurement of observer agreement for categorical data. *Biometrics*. 1977; 33:159–174. [PubMed: 843571]
17. Fox MP, Lash TL, Greenland S. A method to automate probabilistic sensitivity analyses of misclassified binary variables. *International journal of epidemiology*. 2005; 34(6):1370–1376. [PubMed: 16172102]
18. Lash, TL.; Fox, MP.; Fink, AK. *Applying Quantitative Bias Analysis to Epidemiologic Data*. New York: Springer; 2009.
19. Orsini N, Bellocco R, Bottai M, Wolk A, Greenland S. A tool for deterministic and probabilistic sensitivity analysis of epidemiologic studies. *Stata Journal*. 2008; 8:29–48.
20. 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

Highlights

- We examined the validity of a claims-based algorithm to identify CRC recurrence.
- We determined the extent to which recurrence misclassification affects survival.
- Secondary malignancy codes yielded moderate sensitivity and high specificity.
- Identifying recurrence in CRC patients using claims data is feasible.
- Addressing misclassification is important when using claims-based algorithms.

Table 1

Patient Characteristics used in estimating the accuracy of claims data to identify recurrence.

	Recurrence [*]	
	Yes (N=32) n(%)	No (N=142) n(%)
Age, years Mean (std dev)	74.8 (6.3)	73.9 (6.4)
Gender		
Male	16 (50.0)	71 (50.0)
Female	16 (50.0)	71 (50.0)
Race/ethnicity		
White	26 (81.2)	113 (79.6)
Black/Other	6 (18.8)	29 (20.4)
ACE-27 Comorbidity Score		
0	3 (13.0)	22 (19.5)
1	11 (47.8)	42 (37.2)
2	6 (26.1)	34 (30.1)
3+	3 (13.0)	15 (13.3)
Missing	9 (28.1)	29 (20.4)
Primary tumor location		
Colon	21 (65.6)	91 (64.1)
Rectal	11 (34.4)	51 (35.9)
Stage at Diagnosis ^{**}		
Localized	6 (18.8)	68 (47.9)
Regional by Direct Extension	6 (18.8)	26 (18.3)
Regional by lymph nodes only	8 (25.0)	22 (15.5)
Regional by lymph nodes and direct extension	12 (37.5)	26 (18.3)
Chemotherapy		
Any Chemotherapy	18 (56.2)	59 (41.5)
None, not planned	11 (34.4)	73 (51.4)
Recommended, not given/Refused	3 (9.4)	10 (7.0)
Radiation Therapy		
Any	8 (25.0)	37 (26.1)
None	24 (75.0)	105 (73.9)
Vital Status ^{**}		
Died	17 (53.1)	24 (16.9)
Alive	15 (46.9)	118 (83.1)

* Recurrence based on medical record review;

**
Chi-square p-value <0.05

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Characteristics of the claims-based algorithm to identify recurrence.

Algorithm	Sensitivity (95% CI)	Specificity (95% CI)	Predictive value positive (95% CI)	Negative predictive value (95% CI)
Secondary malignancy codes 197, 197.0-197.4, 197.6-197.8, 198.0-198.8, 198.82, 198.89	81.3 (63.0-92.1)	99.3 (95.6-100.0)	96.3 (79.1-99.8)	95.9 (90.9-98.3)
Chemotherapy or radiation 8 months or more post-surgery	6.0 (1.1-22.2)	100 (96.7-100)	100 (19.8-100)	82.6 (75.9-87.7)
Chemotherapy 16 months or more post-surgery or radiation 12 months or more post-surgery	12.5 (4.1-29.9)	98.6 (94.5-99.8)	66.7 (24.1-94.0)	83.3 (76.6-88.5)

CI: confidence interval.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3

Corrected hazard ratios for the association of recurrence and all-cause death using a range of sensitivities (keeping specificity constant at 98.0% (min: 95.0, mode 2: 99.0, max: 100.0)).

Mode 1 sensitivity (min, mode 2, max)	Corrected hazard ratio (2.5% - 97.5%)
75.0 (60.0, 88.0, 93.0)	3.47 (3.06 – 4.41)
80.0 (65.0, 93.0, 98.0)	3.39 (3.00 – 4.00)
85.0 (70.0, 98.0, 100.0)	3.32 (2.97 – 3.91)
90.0 (75.0, 100.0, 100.0)	3.28 (2.95 – 3.84)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript