Brief communication

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

Anjali D. Deshpande PhD, MPH [a,*], Mario Schootman PhD [b], Allese Mayer MPH [b]

[a] Division of General Medical Sciences, Department of Medicine, School of Medicine, Washington University in St. Louis, St. Louis, MO
[b] Department of Epidemiology, College for Public Health and Social Justice, Saint Louis University, St. Louis, MO

ARTICLE INFO

ABSTRACT

Purpose: To examine the validity of claims data to identify colorectal cancer (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 Surveillance, Epidemiology, and End Results—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% confidence interval: 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% confidence interval: 3.06—4.14) more likely to die than those without recurrence.
Conclusions: Identifying recurrence in CRC patients using claims data is feasible with moderate sensitivity and high specificity. Future studies can use this algorithm with Surveillance, Epidemiology, and End Results—Medicare data to study treatment patterns and outcomes of CRC patients with recurrence.

Background

Colorectal cancer (CRC) is the third most common cancer in the United States. 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 the following two components: (1) accuracy of claims data relative to medical records to identify recurrence after 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.

Accuracy of claims data

Data sources and abstraction

We used the following 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, *International Classification of Diseases, Ninth Revision, Clinical Modification* [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 after 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 et al. [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 two possible treatment algorithms of recurrence as follows: (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 ICD-9 and Healthcare Common Procedure Coding System (HCPCS) codes as described by Warren et al. [14].

Statistical analysis

The ODS data, enhanced by medical record abstraction, were 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 because of inconsistent coding.

We calculated sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and associated 95% confidence intervals (CIs) 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 as follows: 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) [15]. Because kappa is affected by prevalence (i.e., recurrence), we also calculated the prevalence bias–adjusted kappa.

Misclassified recurrence and survival

We obtained data from an existing linkage of 2000 to 2005 SEER program data of 12 registries with 1999 to 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) [16] 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 et al. [17]. 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, StataCorp LP, College Station, TX) to adjust the observed hazard ratio for misclassification bias [18].

Results

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