Short Communication

Cancer Epidemiology, Biomarkers & Prevention

Validation of a Coding Algorithm to Identify Bladder Cancer and Distinguish Stage in an Electronic Medical Records Database

Ronac Mamtani[1,2], Kevin Haynes[2], Ben Boursi[2,3], Frank I. Scott[2], David S. Goldberg[2], Stephen M. Keefe[1], David J. Vaughn[1], S. Bruce Malkowicz[1], and James D. Lewis[2]

Abstract

Studies on outcomes in bladder cancer rely on accurate methods to identify patients with bladder cancer and differentiate bladder cancer stage. Medical record and administrative databases are increasingly used to study cancer incidence, but few have distinguished cancer stage, and none have focused on bladder cancer. In this study, we used data from The UK Health Improvement Network (THIN) to identify patients with bladder cancer using at least one diagnostic code for bladder cancer, and distinguish muscle-invasive from non-invasive disease using a subsequent code for cystectomy. Algorithms were validated against a gold standard of physician-completed questionnaires, pathology reports, and consultant letters. Algorithm performance was evaluated by measuring positive predictive value (PPV) and corresponding 95% confidence interval (CI). Among all patients coded with bladder cancer ($n = 194$), PPV for any bladder cancer was 99.5% (95% CI, 97.2–99.9). PPV for incident bladder cancer was 93.8% (95% CI, 89.4–96.7). PPV for muscle-invasive bladder cancer was 70.1% (95% CI, 59.4–79.5) in patients with cystectomy ($n = 95$) and 83.9% (95% CI, 66.3–94.5) in those with cystectomy plus additional codes for metastases and death ($n = 31$). Using our codes for bladder cancer, the age- and sex-standardized incidence rate (SIR) of bladder cancer in THIN approximated that measured by cancer registries (SIR within 20%), suggesting that sensitivity was high as well. THIN is a valid and novel database for the study of bladder cancer. Our algorithm can be used to examine the epidemiology of muscle-invasive bladder cancer or outcomes following cystectomy for patients with muscle invasion. *Cancer Epidemiol Biomarkers Prev; 24(1); 303–7. ©2014 AACR.*

Introduction

Bladder cancer is the sixth most common cancer in the United States and the ninth most common worldwide (1, 2). Most bladder cancers are non-muscle invasive (i.e., superficial) at diagnosis, yet muscle-invasive tumors are the major cause of morbidity and mortality from this disease with 5-year survival rates of only 40% to 60% (3). Bladder cancer stage at initial diagnosis is therefore an important predictor of disease outcome.

Research on bladder cancer epidemiology and outcomes has been limited by a lack of large-scale datasets with valid information on cancer stage and recurrence. Cancer registry programs [e.g., Surveillance, Epidemiology, and End Results (SEER)] collect data on tumor stage, but not tumor recurrence (http://seer.cancer.gov/about/overview.html). Administrative claims data have frequently been used to evaluate the accuracy of incident cancers (4–9), but few have attempted to distinguish cancer stage (10–13), and none have focused on bladder cancer.

Electronic medical record (EMR) databases are increasingly used to study cancer incidence and outcomes. EMR data contain important exposure information lacking in most cancer registry and administrative data, such as medical conditions, medications, smoking, and body mass index. In this study, we evaluated the validity of an EMR to (i) accurately identify patients with a bladder cancer diagnosis and (ii) distinguish muscle-invasive from non-invasive bladder cancer against the gold-standard chart review.

Materials and Methods

Data source

We developed algorithms using data from The Health Improvement Network (THIN), a primary care medical records database that is representative of the broader UK population (14). The database currently contains records of over 11 million patients. Data available in THIN include demographic information, medical diagnoses including surgical procedures, lifestyle characteristics such as smoking status, and other clinical measurements recorded by general practitioners (GPs), such as body mass index. Medical diagnoses within the database are recorded using Read codes, the standard primary care classification system in the United Kingdom (15). The accuracy and completeness of THIN is well documented for several chronic diseases, including some cancers (16–18). In addition to the EMR, THIN provides researchers with access to written records, including surgical pathology, operative notes, consultant reports, and death certificates. These

[1]Abramson Cancer Center, University of Pennsylvania, Philadelphia, Pennsylvania. [2]Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania, Philadelphia, Pennsylvania. [3]Tel Aviv Sourasky Medical Center, Tel Aviv, Israel.

Note: Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

Corresponding Author: Ronac Mamtani, University of Pennsylvania, 16 Penn Tower, 3400 Spruce Street, Philadelphia, PA 19104. Phone: 215-615-1607; Fax: 215-614-0456; E-mail: ronac.mamtani@uphs.upenn.edu

doi: 10.1158/1055-9965.EPI-14-0677

documents can be supplemented with surveys administrated to the GPs, as was done in this study.

Study design and population

We conducted a cross-sectional study among patients in THIN ≥21 years of age with at least 6 months of follow-up preceding a first diagnostic code for bladder cancer occurring after 2001. We excluded patients with a first bladder cancer diagnosis before or within 6 months of a patient's registration with the GP to avoid misclassification of prevalent bladder cancer as incident bladder cancer. To ensure complete recording of cancer diagnoses in THIN, we excluded patients with a bladder cancer diagnosis occurring before 2001 *a priori* (17).

Data collection and primary outcome definition

Using stratified random sampling (Fig. 1), we surveyed GPs caring for patients in THIN with codes predictive of any bladder cancer (≥1 bladder cancer code with or without cystectomy code) and muscle-invasive bladder cancer (≥1 bladder cancer code with subsequent cystectomy code). We used cystectomy as a marker for muscle invasion, given that cystectomy remains the standard therapy for patients with muscle-invasive tumor (19). The mailed questionnaire (Supplementary Fig. S1) asked the GP to confirm the bladder cancer diagnosis, provide the date and stage at first diagnosis, indicate whether the subject underwent radical cystectomy, and if so, whether the patient developed recurrence (i.e., metastases) after surgery. The GP was also asked to provide copies of all pathology reports, consultant letters, and death certificates relevant to the diagnosis. For each outcome, recorded diagnoses in the electronic record were compared with the data from both physician-completed questionnaires and medical reports as the gold standard.

Statistical analysis

Algorithm performance was evaluated by measuring its positive predictive value (PPV) and corresponding 95% confidence interval (CI). We focused on PPV because if this parameter is sufficiently high, researchers can have confidence that the algorithm will identify subjects with high probabilities of having a true bladder cancer event. PPV for a bladder cancer diagnosis was calculated as the proportion of total patients with coded bladder cancer documented as having true bladder cancer by GP questionnaire or chart review as the gold standard. PPV for muscle invasion was calculated among patients coded for bladder cancer and subsequent cystectomy. Similar methods were used to calculate PPV for muscle invasion among separate cystectomy subgroups (cystectomy without codes for metastases or death, cystectomy with codes for metastases only, and cystectomy with codes for metastases and death). We also measured PPV for an incident bladder cancer event. We determined whether the bladder cancer was incident or recurrent from GP questionnaires, which captured the date of the patient's first diagnosis with bladder cancer. We

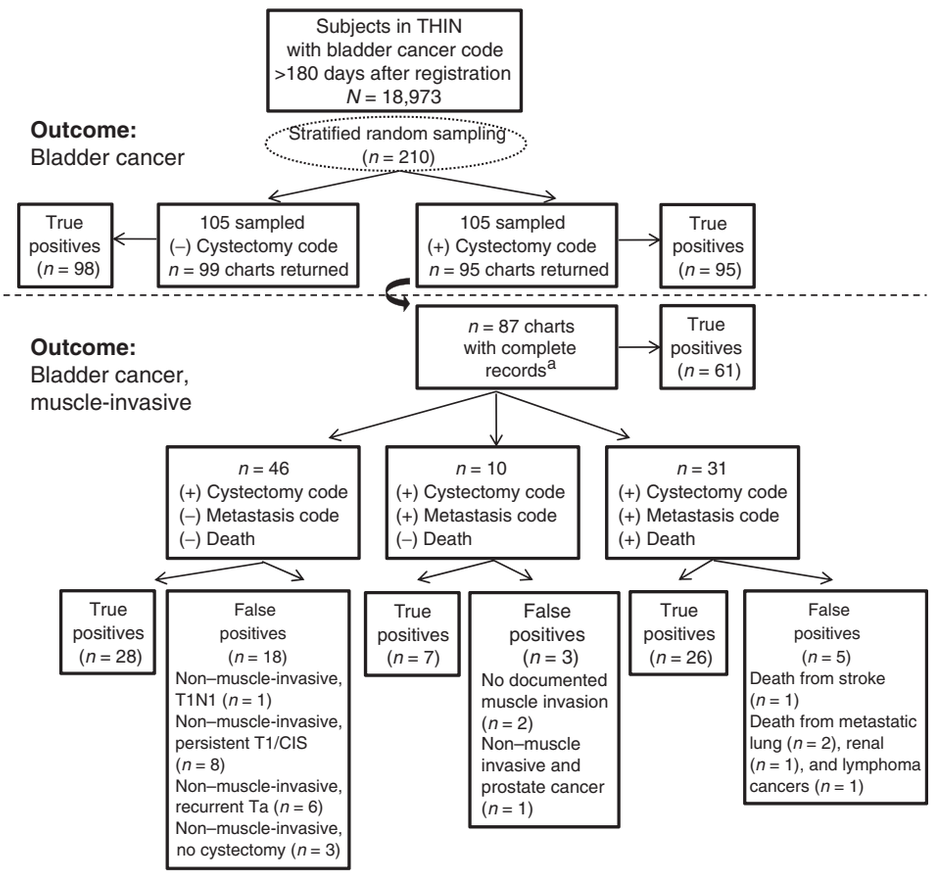

Figure 1.
Algorithms used to identify subjects with any bladder cancer and muscle-invasive bladder cancer. True positives indicate that the criterion was met and confirmed against the gold-standard chart review. False positives indicate that the criterion was met but not confirmed. Metastasis code, at least one secondary neoplasm code; lymph node, liver, or bone metastasis code; or bladder cancer code occurring >180 days after cystectomy code. [a]Eighteen patients were excluded due to unreturned charts ($n = 10$) or incomplete records ($n = 8$).

considered the THIN record to have accurately identified an incident bladder cancer diagnosis if the date of first diagnosis recorded in THIN was within 30 days of the date of diagnosis recorded in the gold-standard data.

Next, we tested whether using our codes for bladder cancer, the age- and sex-SIRs of bladder cancer in THIN are comparable with those in the UK cancer registry. Finally, among subjects with a coded diagnosis of bladder cancer, we estimated the proportion of those with confirmed muscle invasion that were captured by the algorithm using sample weights to account for the sampling fractions used in this study.

We estimated that a sample of 100 patients with cystectomy and 100 patients without cystectomy would allow determination of the PPV within a confidence interval width of ±8%, respectively. We oversampled and requested 105 patients each (N = 210), assuming that 5% of the charts would not be returned. The diagnostic codes used for the algorithms in this study can be found in Supplementary Table S1.

STATA version 12.0 was used for all statistical analyses (Stata Corp).

Results

Questionnaires and other written records (i.e., pathology and consultant reports) were returned from 92% (n = 194) of the 210 subjects sampled. Table 1 describes the demographics of each cohort. Most subjects were male (75%) and smokers (78%), with median ages at diagnosis ranging from 67 to 70. Those for whom the surveys were not returned were less likely to be documented smokers in the EMR.

Agreement between physician-survey and medical chart documentation of a bladder cancer event was 100%. Table 2 demonstrates the PPVs for bladder cancer and muscle-invasive bladder cancer in all patients, patients with cystectomy, and subsets of cystectomy. Among all patients with codes predictive of bladder cancer (N = 194), 193 were correctly classified corresponding to 99.5% PPV (95% CI, 97.2–99.9) and 182 were incident diagnoses (PPV 93.8%; 95% CI, 89.4–96.7; Supplementary Table S2). PPVs for bladder cancer were similar in patients with or without cystectomy (PPVs = 99% vs. 100%). The PPV for muscle invasion in patients with cystectomy (n = 95) was 70.1% (95% CI, 59.4–79.5), with the highest PPV observed in cystectomy patients with additional codes for metastases and death (n = 31; PPV = 83.9%; 95% CI, 66.3–94.5) and the lowest PPV in cystectomy patients without the additional codes for metastases or death (n = 46; PPV = 61.0%; 95% CI, 45.4–74.9). In the full THIN population, there were 1,347 patients with codes suggestive of muscle-invasive bladder cancer. The weighted PPV for muscle invasion after accounting for the proportional representation of these subsets in the full THIN population was 74.8% (95% CI, 62.6–87.4).

Of patients meeting the algorithm for muscle invasion and not confirmed to have muscle-invasive bladder cancer (n = 26 false positives; Fig. 1), most patients had persistent or recurrent non–muscle-invasive bladder cancer (n = 14) requiring cystectomy. Because non–muscle-invasive bladder cancers frequently recur and require repeat transurethral resection of the bladder (TURB) over time, we examined whether more frequent codes for TURB recorded before cystectomy and increased time from the first recorded bladder cancer diagnosis to the date of cystectomy improves algorithm performance. However, including these data did not improve PPV for muscle invasion (data not shown).

To further assess the reliability of our bladder cancer diagnostic codes, we calculated the annual SIR of bladder cancer in THIN using age- and sex-specific rates from the UK cancer registry data. SIRs for bladder cancer were within 20% of unity using data after 2001 (Supplementary Fig. S2). Finally, the proportion of patients that met the primary algorithm for muscle invasion among those with confirmed invasive disease was 40.0% (95% CI, 12.1–73.8) after accounting for sampling weights (Supplementary Table S3).

Discussion

This study examined the accuracy of diagnostic codes in a UK medical records database to identify patients with bladder cancer and distinguish muscle-invasive from superficial disease. An algorithm using at least one diagnostic code for bladder cancer had high PPV (99%) for true bladder cancer and likely high sensitivity based on the computed SIRs. A subsequent code for cystectomy had modest PPV (70%) for muscle invasion. PPV increased to 84% when patients had additional codes for metastasis and death after a code for cystectomy, suggesting that the algorithm can be particularly useful in studies examining risk factors for muscle-invasive bladder cancer.

Methods to identify patients with bladder cancer and differentiate bladder cancer stage using medical record or administrative data have not been developed. Among patients with bladder cancer, muscle invasion is a key predictor of survival (19). Yet, risk factors for the development of muscle-invasive versus superficial tumor are unknown. The simple algorithms developed in this study can therefore be used to facilitate future studies to examine the epidemiology of muscle-invasive bladder cancer or outcomes following cystectomy for patients with muscle invasion.

When developing algorithms, researchers must prioritize different measures of test accuracy (i.e., sensitivity, specificity, PPV, and negative predictive value), depending on the goals of the study (20). Epidemiologists often prioritize PPV over sensitivity so as to minimize misclassification. The PPV for any bladder cancer was extremely high, nearly 100%. Although we could not directly measure sensitivity for any bladder cancer, the incidence of bladder cancer in THIN approximated that measured by cancer registries, suggesting that sensitivity was high as well. Results from our study suggest that 25% of patients identified as having muscle-invasive bladder cancer using exclusively diagnostic (bladder cancer) and procedure (cystectomy) codes will be false positives (weighted PPV 75%) and 60% of all patients with muscle invasion may not be readily identifiable as having muscle invasion using our algorithm. Importantly, most bladder cancers misclassified as muscle invasive were non–muscle-invasive

Table 1. Characteristics of study participants

Characteristic	Patients without cystectomy (n = 99)	Patients with cystectomy (n = 95)	Surveys not returned (n = 16)
Age at diagnosis, y, median (IQR)	69.9 (65.7–78.3)	66.9 (60.8–71.3)	68.3 (63.3–74.5)
Male sex, n (%)	75 (75.8)	73 (76.8)	10 (62.5)
Smoking, n (% ever)	78 (78.8)	78 (82.1)	9 (56.3)

Table 2. Algorithm performance

Patient group	Number sampled	Number in THIN	PPV for bladder cancer, 95% CI	PPV for muscle-invasive bladder cancer, 95% CI
All patients	194	18, 973	99.5 (97.2–99.9)	N/A
Patients without cystectomy	99	17,626	99.0 (94.5–99.9)	N/A
Patients with cystectomy	95[a]	1,347	100.0 (96.2–100.0)	70.1 (59.4–79.5)
Cystectomy without metastases or death	46	485	100.0 (92.3–100.0)	61.0 (45.4–74.9)
Cystectomy with metastases, alive	10	78	100.0 (69.2–100.0)	70.0 (34.7–93.3)
Cystectomy with metastases, dead	31	784	100.0 (88.8–100.0)	83.9 (66.3–94.5)
Weighted calculation[b]	—	—	—	74.8 (62.6–87.4)

[a]Eight patients were excluded due to incomplete records.

[b]PPV for muscle-invasive bladder cancer using the proportional representation of cystectomy subsets in the full THIN population [(0.61 × 485/1347) + (0.70 × 78/1347) + (0.839×784/1347)].

cancers and considered to be high risk for progression and preemptively treated with cystectomy (as is the case in some centers; ref. 21). Examining other indicators for muscle invasion, in addition to cystectomy (the standard therapy for patients with muscle invasion), such as codes predictive for metastasis or death, improved PPV but would be expected to reduce sensitivity.

The degree of bias that may result from misclassifying muscle-invasive tumors as non-invasive would depend on the study design. For example, we estimated bias-adjusted exposure odds ratios for muscle invasion in a case–control study. A 30% non-differential case misclassification rate (corresponding to 70% PPV) resulted in only a modest attenuation of the magnitude of the measured OR (15%–18%), assuming exposure prevalence of 10% to 30% and a true OR of 1.5. Investigators can use these data to weigh the cost versus the utility of requesting medical records to validate muscle invasion.

Strengths of this study included the high response rate (>90%), stratified random sampling among all enumerated subjects, and rigorous validation of bladder cancer diagnoses against gold-standard chart review. In addition, the PPVs for an incident bladder cancer were exceptionally high, suggesting that exclusion of patients with a first diagnosis of bladder cancer before 6 months within registration is sufficient to accurately identify newly diagnosed bladder cancer. We also demonstrated the completeness of bladder cancer recording in THIN after 2001 using UK cancer registry data as the standard of reference, as has been reported with other solid tumors (17). Finally, the results of this study are likely generalizable to the Clinical Practice Research Datalink (CPRD), a related UK database for which there is overlap in practices with THIN and that uses the same EMR software for data collection. Further studies are required to test whether our algorithm is generalizable to U.S. datasets using ICD-9 codes.

In conclusion, we demonstrate that THIN is a valid database for the study of bladder cancer. Our algorithm can be used to identify patients with bladder cancer with high PPV and likely high sensitivity. The PPV for muscle-invasive disease is not quite as high overall, but is quite good in subsets with codes for metastatic disease or death.

Disclosure of Potential Conflicts of Interest

J.D. Lewis reports receiving a commercial research grant from Takeda and is a consultant/advisory board member for Takeda and Janssen. No potential conflicts of interest were disclosed by the other authors.

Authors' Contributions

Conception and design: R. Mamtani, K. Haynes, F.I. Scott, J.D. Lewis
Development of methodology: R. Mamtani, K. Haynes, B. Boursi
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): K. Haynes, S.B. Malkowicz
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): R. Mamtani, K. Haynes, F.I. Scott, D.S. Goldberg, S.B. Malkowicz, J.D. Lewis
Writing, review, and/or revision of the manuscript: R. Mamtani, K. Haynes, B. Boursi, F.I. Scott, D.S. Goldberg, S.M. Keefe, D.J. Vaughn, S.B. Malkowicz, J.D. Lewis
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): R. Mamtani, K. Haynes
Study supervision: K. Haynes, J.D. Lewis

Grant Support

This research was supported by the NIH (grant number K12 CA 076931 to R. Mamtani, K08-DK095951-02 to F.I. Scott, K08-DK098272-01 to D.S. Goldberg, UL1-RR024134 to K. Haynes and J.D. Lewis, and K24-DK078228 to J.D. Lewis) and the Conquer Cancer Foundation of the American Society of Clinical Oncology (Young Investigator Award to R. Mamtani).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received June 16, 2014; revised October 27, 2014; accepted October 31, 2014; published OnlineFirst November 11, 2014.

References

1. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2013. CA Cancer J Clin 2013;63:11–30.
2. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin 2011;61:69–90.
3. Stein JP, Lieskovsky G, Cote R, Groshen S, Feng AC, Boyd S, et al. Radical cystectomy in the treatment of invasive bladder cancer: long-term results in 1,054 patients. J Clin Oncol 2001;19:666–75.
4. Setoguchi S, Solomon DH, Glynn RJ, Cook EF, Levin R, Schneeweiss S. Agreement of diagnosis and its date for hematologic malignancies and solid tumors between Medicare claims and cancer registry data. Cancer Causes Control 2007;18:561–9.
5. Baldi I, Vicari P, Di Cuonzo D, Zanetti R, Pagano E, Rosato R, et al. A high positive predictive value algorithm using hospital administrative data identified incident cancer cases. J Clin Epidemiol 2008;61:373–9.
6. Cooper GS, Yuan Z, Stange KC, Dennis LK, Amini SB, Rimm AA. The sensitivity of Medicare claims data for case ascertainment of six common cancers. Med Care 1999;37:436–44.
7. Nattinger AB, Laud PW, Bajorunaite R, Sparapani RA, Freeman JL. An algorithm for the use of Medicare claims data to identify women with incident breast cancer. Health Serv Res 2004;39:1733–49.
8. Ramsey SD, Scoggins JF, Blough DK, McDermott CL, Reyes CM. Sensitivity of administrative claims to identify incident cases of lung cancer: a comparison of 3 health plans. J Manag Care Pharm 2009;15:659–68.
9. Goldberg DS, Lewis JD, Halpern SD, Weiner MG, Lo Re V 3rd. Validation of a coding algorithm to identify patients with hepatocellular

carcinoma in an administrative database. Pharmacoepidemiol Drug Saf 2013;22:103–7.
10. Nordstrom BL, Whyte JL, Stolar M, Mercaldi C, Kallich JD. Identification of metastatic cancer in claims data. Pharmacoepidemiol Drug Saf 2012;21 Suppl 2:21–8.
11. Smith GL, Shih YC, Giordano SH, Smith BD, Buchholz TA. A method to predict breast cancer stage using Medicare claims. Epidemiol Perspect Innov 2010;7:1.
12. Chubak J, Yu O, Pocobelli G, Lamerato L, Webster J, Prout MN, et al. Administrative data algorithms to identify second breast cancer events following early-stage invasive breast cancer. J Natl Cancer Inst 2012;104:931–40.
13. Whyte JL, Engel-Nitz NM, Teitelbaum A, Gomez Rey G, Kallich JD. An evaluation of algorithms for identifying metastatic breast, lung, or colorectal cancer in administrative claims data. Med Care 2013 Mar 21. [Epub ahead of print].
14. Blak BT, Thompson M, Dattani H, Bourke A. Generalisability of the health improvement network (THIN) database: demographics, chronic disease prevalence and mortality rates. Inform Prim Care 2011;19:251–5.
15. Chisholm J. The Read clinical classification. BMJ 1990;300:1092.
16. Lewis JD, Schinnar R, Bilker WB, Wang X, Strom BL. Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. Pharmacoepidemiol Drug Saf 2007;16:393–401.
17. Haynes K, Forde KA, Schinnar R, Wong P, Strom BL, Lewis JD. Cancer incidence in the health improvement network. Pharmacoepidemiol Drug Saf 2009;18:730–6.
18. Iyen-Omofoman B, Hubbard RB, Smith CJ, Sparks E, Bradley E, Bourke A, et al. The distribution of lung cancer across sectors of society in the United Kingdom: a study using national primary care data. BMC Public Health 2011;11:857.
19. McDougal W, Shipley W, Kaufman D, Dahl D, Michaelson M, Zietman A. DeVita, Hellman, and Rosenberg's Cancer: principles and practice of oncology. In: DeVita V, Lawrence T, Rosenberg S, DePinho R, Weinberg R, editors. Cancer of the bladder, ureter, and renal pelvis. 9th ed. Philadelphia, PA: Lippencott Williams & Wilkins; 2011. p. 1192–204.
20. Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. J Clin Epidemiol 2012;65:343–9.e2.
21. Herr HW, Sogani PC. Does early cystectomy improve the survival of patients with high risk superficial bladder tumors? J Urol 2001;166:1296–9.

Validation of a Coding Algorithm to Identify Bladder Cancer and Distinguish Stage in an Electronic Medical Records Database

Ronac Mamtani, Kevin Haynes, Ben Boursi, et al.

Cancer Epidemiol Biomarkers Prev 2015;24:303-307. Published OnlineFirst November 11, 2014.