Machine Learning Methods to Identify Missed Cases of Bladder Cancer in Population-Based Registries

Anne-Michelle Noone, MS[1]; Clara J. K. Lam, PhD, MPH[1]; Angela B. Smith, MD, MS[2,3]; Matthew E. Nielsen, MD, MS[2,3]; Eric Boyd, BS[4]; Angela B. Mariotto, PhD[1]; and Mousumi Banerjee, PhD[5]

abstract

PURPOSE Population-based cancer incidence rates of bladder cancer may be underestimated. Accurate estimates are needed for understanding the burden of bladder cancer in the United States. We developed and evaluated the feasibility of a machine learning–based classifier to identify bladder cancer cases missed by cancer registries, and estimated the rate of bladder cancer cases potentially missed.

METHODS Data were from population-based cohort of 37,940 bladder cancer cases 65 years of age and older in the SEER cancer registries linked with Medicare claims (2007-2013). Cases with other urologic cancers, abdominal cancers, and unrelated cancers were included as control groups. A cohort of cancer-free controls was also selected using the Medicare 5% random sample. We used five supervised machine learning methods: classification and regression trees, random forest, logic regression, support vector machines, and logistic regression, for predicting bladder cancer.

RESULTS Registry linkages yielded 37,940 bladder cancer cases and 766,303 cancer-free controls. Using health insurance claims, classification and regression trees distinguished bladder cancer cases from noncancer controls with very high accuracy (95%). Bacille Calmette-Guerin, cystectomy, and mitomycin were the most important predictors for identifying bladder cancer. From 2007 to 2013, we estimated that up to 3,300 bladder cancer cases in the United States may have been missed by the SEER registries. This would result in an average of 3.5% increase in the reported incidence rate.

CONCLUSION SEER cancer registries may potentially miss bladder cancer cases during routine reporting. These missed cases can be identified leveraging Medicare claims and data analytics, leading to more accurate estimates of bladder cancer incidence.

JCO Clin Cancer Inform 5:641-653. © 2021 by American Society of Clinical Oncology

ASSOCIATED CONTENT

Appendix

Data Supplement

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on April 26, 2021 and published at ascopubs.org/journal/cci on June 7, 2021: DOI https://doi.org/10.1200/CCI.20.00170

BACKGROUND

Cancer surveillance relies on a comprehensive system to collect information on newly diagnosed patients with cancer. This information is critical to accurately estimate cancer incidence and survival. In turn, these data provide a basis for public health research to understand and work toward reducing the cancer burden in the population. Data integrity is dependent upon cancer surveillance meeting high standards of completeness.

The North American Association of Central Cancer Registries estimates the extent to which all incident cases are reported to the registry using the incidence-mortality ratio.[1-3] Assumptions of this method may not be met since external factors may influence cancer incidence such as trends in cancer risk factors or screening; therefore, this ratio may not fully capture registry case completeness.[4] Furthermore, cases were historically diagnosed and treated in the hospital setting and easily identified by cancer registrars. Over time, cancer diagnosis and treatment have evolved substantially, often now in outpatient settings, making measures of case completeness even more critical. Particularly, bladder cancer is often diagnosed and treated in urology offices and may not get reported routinely to cancer registries.[5,6] Therefore, they are not included in population-based incidence rate estimates, potentially resulting in underestimation of bladder cancer incidence.

Bladder cancer diagnosis and treatment cascade is unique compared with other cancers. Specifically, a diagnostic and initial therapeutic procedure is often a transurethral resection of bladder tumor (TURBT).[7] The goal of this procedure is to make the correct diagnosis and remove visible lesions. Another common treatment is intravesical Bacille Calmette-Guerin (BCG) after surgery. If the tumor has invaded the

CONTEXT

Key Objective

To develop and evaluate the feasibility of a machine learning–based classifier to identify bladder cancer cases potentially missed by cancer registries.

Knowledge Generated

A classification tree was able to identify bladder cancer cases versus noncancer controls with very high accuracy using treatment and comorbidity information from medical claims. Common treatments for bladder cancer including Bacille Calmette-Guerin, cystectomy, and mitomycin were important predictors for identifying bladder cancer cases. We estimated that the incidence rate of bladder cancer reported by cancer registries is likely to be underestimated by 3.5%.

Relevance

Cancer registries may not record all cases of bladder cancer primarily because of diagnosis and treatment outside of hospital settings. Machine learning–based classifiers, such as a classification tree, may accurately identify these unrecorded cases. This would lead to more accurate reporting of bladder cancer incidence rates.

muscle, then either a radical or partial cystectomy is standard. Patients may also receive combinations of radiation and chemotherapy depending on tumor stage.[7,8] Because many of these procedures are specific to bladder cancer, there is an opportunity to develop an algorithm using procedure codes to identify unreported cases.

Machine learning methods are powerful statistical techniques for developing classification tools. In contrast to tools based on clinical knowledge of disease and treatment, these methods choose the best algorithm that results in the lowest misclassification error. Machine learning methods can easily handle large amounts of data and many predictor variables. They are well suited to identify nonlinear relationships including interactions or Boolean combinations of variables that may not be known a priori. To our knowledge, there is no published research using these techniques to identify unreported cases of cancer. We used five machine learning methods, specifically logistic regression, classification and regression trees (CART), random forest, support vector machines (SVM), and logic regression, to build a classifier, and compared these based on predictive accuracy.

The primary objective was to develop an algorithm for cancer surveillance (ie, to detect unreported cases of bladder cancer) that is critical for estimating bladder cancer incidence in the United States. Furthermore, these results may also highlight whether claims data could be used by registries to ascertain unreported bladder cancers.

METHODS

Data Sources

The SEER program of the National Cancer Institute is a system of 18 population-based cancer registries that covers 35% of the US population from geographically defined areas. Individuals in the SEER data eligible for Medicare have been matched to their Medicare claims to create the linked SEER-Medicare data. These linked data contain longitudinal claims with codes for medical services and diagnoses associated with services and dates.[9] Specifically, individuals in the SEER data were matched to Medicare's master enrollment file maintained by the Centers for Medicare and Medicaid Services. Ninety-four percent of those reported to SEER 65 years of age or older have been linked to their Medicare claims. The linked database also contains a random sample of Medicare beneficiaries who do not have cancer. This cancer-free group is a random 5% sample of Medicare beneficiaries residing in the SEER areas without a cancer recorded in a SEER registry. The Medicare claims in both the SEER-Medicare data and the 5% random sample include hospital care (part A) as well as physician and outpatient services (part B). Part B requires that a beneficiary pay a premium and includes International Classification of Diseases (ICD)-9 and ICD-9 diagnosis codes and Health Care Procedure Healthcare Common Procedure Coding System codes for treatments. Similar claims data are available for the 5% noncancer sample.

Study Sample

We included all individuals in the SEER-Medicare data who were diagnosed with nonmetastatic bladder cancer from 2007 to 2013. We also constructed several control groups to distinguish bladder cancer from similar cancers and those without cancer. The first set was patients diagnosed with other urologic cancers (ie, kidney and renal pelvis, ureter, and other urinary organs); the second was a set of other abdominal cancers (ie, stomach, liver, pancreas, colon, rectum, and gallbladder); and the third was a set of unrelated cancers (eg, female breast and prostate). Cancer types were defined by the SEER ICD-O-3 codes.[10] We included in our study all consecutive individuals who were at least 65 years of age, had continuous part A, B, and fee-for-service coverage, and not enrolled in a Health Maintenance Organization during that year. Individuals could have more than one cancer recorded by SEER. If this was the case, the first primary cancer was used to determine if they were selected into the bladder cancer group or one of the other cancer groups. We also selected controls without a cancer diagnosis from the 5% noncancer sample that comprised a fourth control group. Ten cancer-free controls were randomly selected for each bladder cancer case matched by birth year. Thus, we constructed four data sets using the four different control groups and each was randomly split into 2/3 for training and 1/3 for testing.

Identification of Medical Conditions and Cancer Treatment

An individual was considered to have a specified medical condition if they had a claim with a diagnosis code for that condition in the year before cancer diagnosis. For the cancer-free controls, conditions that occurred in the year before diagnosis of their matched case were ascertained. Medical conditions included individual comorbidities as well as chronic obstructive pulmonary disease and history of smoking (Table 1). Information on cancer treatment and medical procedures was identified using Medicare claims. An individual was considered to have treatment if at least one Medicare claim included a code for the specific treatment or procedure within the first year of cancer diagnosis. Codes used to identify treatment are listed in Appendix Tables A1 and A2.

Statistical Analyses

We used five supervised machine learning methods, namely, CART, random forest, logic regression, SVM, and logistic regression, for predicting bladder cancer. We included all variables listed in Table 1 in the models except for TURBT. TURBT was excluded because it is used as a confirmatory procedure for bladder cancer. So, using TURBT to predict bladder cancer would not be clinically useful since patients who underwent TURBT but did not have bladder cancer would result in false positives.

CART is a binary recursive partitioning technique.[11-13] The method begins with all subjects in the top node of the tree. Subjects are passed down the tree with decisions made at each node to split into two daughter nodes until no further splitting is done and a terminal node is reached. Each nonterminal node contains a question on which a split is based. In a classification tree, the covariate space is partitioned recursively in a binary fashion based on homogeneity of the nodes. To prevent overfitting, this tree is pruned using a cost-complexity parameter that imposes a penalty for large trees.[14] The final tree is selected based on a 10-fold cross-validation[11] and is the one that has the lowest misclassification error to predict bladder cancer.

Random forest is an ensemble of unpruned classification or regression trees grown using bootstrap resamples of the data.[13,14] This method overcomes much of the inherent instability with a single CART tree. Here, a tree is grown on bootstrap samples using a random selection of covariates at each step of the tree growing. A patient is classified by each tree and the final classification is the one with majority votes across all trees in the forest. We grew a forest of 500 trees to predict bladder cancer. The Gini index, a measure of variable importance, was also estimated.[14] This measures the impact a single variable has on the error rates of the forest and is used to rank the variables.

Logic regression is an adaptive classification and regression procedure that constructs Boolean combinations of binary predictors.[15,16] For example, a decision to classify a patient in a certain group may be based on rules such as if X1 *and* X2 but *not* X3 are true. Logic regression searches among all possible Boolean combinations of predictors while remaining in the regression modeling framework. The quality of the models is determined by an appropriate score function. In our analysis, we used binomial deviance as the score function to predict bladder cancer, and a stochastic optimization algorithm to search for the Boolean expressions.[15,16]

We also used SVM, which is a nonprobabilistic supervised learning procedure that creates a multidimensional hyperplane to partition the covariate space into two groups allowing for classification.[17,18] SVMs create hyperplanes by maximizing the margin between the nearest data points on either side of the hyperplane based on a cost penalty for each misclassified patient.

Finally, we estimated the predicted probabilities of being a bladder cancer case using logistic regression. All covariates were entered into the model and no model selection was performed. Patients were classified as having bladder cancer if their predicted probability was > 50%.

Models were trained and tested using the split sample: 2/3 were used for training and the remaining 1/3 was used as a test set. Classification performance was evaluated using overall accuracy, sensitivity, and specificity based on the test set. Overall accuracy is the proportion of correct classifications using the classifier, and sensitivity is the proportion of individuals classified as having bladder cancer among those who were reported to SEER data (true disease status). Specificity is the proportion of individuals classified as not having bladder cancer among those who did not have a bladder cancer diagnosis. We also calculated the area under the receiver operating characteristic curve and the F1 score.[19] Analysis was conducted using R software, version 3.5.1.[20] Specifically, we used the packages rpart[21] (CART), randomForest[22] (random forest), e1071[23] (SVM), and LogicReg[24] (logic regression). R code and model parameters are available in Appendix Table A3 and the Data Supplement.

Estimates of Unreported Bladder Cancer

We used the entire noncancer sample from Medicare to estimate the number and rate of potential missed cases of bladder cancer in SEER. The date of the first claim for any treatment shown in Table 1 was used as the index date and with the same inclusion and exclusion criteria stated above. That is, the person must have been at least 65 years of age and resided in a SEER registry area before the index date

TABLE 1. Demographic, Comorbid, and Treatment Characteristics of 766,303 Individuals From SEER-Medicare With Selected Cancers and Without Cancer

	Bladder Cancer		Other Urinary Cancers[a]		Abdominal Cancers[a]		Other Cancers[a]		Cancer-Free Controls	
	n = 37,940		n = 22,389		n = 120,450		n = 206,125		n = 379,399	
Patient Characteristics	No.	%	No.	%	No.	%	No.	%	No.	%
Demographic characteristics										
Age, years										
65-69	7,253	19.1	5,989	26.7	24,135	20.0	64,089	31.1	72,692	19.2
70-74	7,961	21.0	5,456	24.4	24,597	20.4	55,274	26.8	79,318	20.9
75-79	8,045	21.2	4,462	19.9	23,869	19.8	40,329	19.6	80,567	21.2
80-84	7,444	19.6	3,550	15.9	22,824	18.9	26,094	12.7	74,283	19.6
85+	7,237	19.1	2,932	13.1	25,025	20.8	20,339	9.9	72,539	19.1
Sex										
Male	28,091	74.0	12,942	57.8	57,957	48.1	113,536	55.1	152,286	40.1
Female	9,849	26.0	9,447	42.2	62,493	51.9	92,589	44.9	227,113	59.9
Race										
Non-Hispanic White	33,073	87.2	17,613	78.7	89,798	74.6	160,513	77.9	299,685	79.0
Non-Hispanic Black	1,623	4.3	1,800	8.0	11,463	9.5	20,281	9.8	30,983	8.2
Hispanic	1,506	4.0	1,711	7.6	9,093	7.5	11,946	5.8	13,011	3.4
API	1,270	3.3	1,151	5.1	9,600	8.0	9,433	4.6	22,010	5.8
Other race	468	1.2	114	0.5	496	0.4	3,952	1.9	13,710	3.6
Diagnosis year										
2007	5,640	14.9	3,239	14.5	18,884	15.7	33,350	16.2	56,399	14.9
2008	5,595	14.7	3,255	14.5	18,375	15.3	31,519	15.3	55,950	14.7
2009	5,286	13.9	3,170	14.2	17,487	14.5	30,012	14.6	52,860	13.9
2010	5,560	14.7	3,136	14.0	16,938	14.1	29,219	14.2	55,600	14.7
2011	5,243	13.8	3,168	14.1	16,611	13.8	29,411	14.3	52,430	13.8
2012	5,398	14.2	3,205	14.3	16,341	13.6	26,527	12.9	53,980	14.2
2013	5,218	13.8	3,216	14.4	15,814	13.1	26,087	12.7	52,180	13.8
Summary stage 2000										
No cancer									379,399	
In situ	18,905	49.8	565	2.5	2,866	2.4	14,632	7.1		
Localized	13,569	35.8	11,466	51.2	37,103	30.8	140,804	68.3		
Regional	2,522	6.6	3,571	15.9	36,050	29.9	29,555	14.3		
Distant	1,506	4.0	3,459	15.4	29,814	24.8	11,515	5.6		
Unknown or not staged[b]	1,438	3.8	3,328	14.9	14,617	12.1	9,619	4.7		
Comorbid conditions (No. and percentage with condition)										
Acute MI	559	1.5	326	1.5	1,899	1.6	1,319	0.6	3,433	0.9
History of MI	1,471	3.9	853	3.8	3,692	3.1	3,445	1.7	5,559	1.5
AIDS	25	0.1	17	0.1	80	0.1	121	0.1	117	0.0
Congestive heart failure	4,319	11.4	2,703	12.1	15,176	12.6	12,295	6.0	25,045	6.6
COPD	6,784	17.9	3,674	16.4	19,185	15.9	20,801	10.1	31,137	8.2
Cerebrovascular disease	3,033	8.0	1,703	7.6	9,342	7.8	10,290	5.0	18,202	4.8
Dementia	1,371	3.6	682	3.0	5,517	4.6	4,580	2.2	13,966	3.7
Diabetes	9,673	25.5	6,531	29.2	35,878	29.8	42,419	20.6	53,558	14.1
Diabetes with complications	2,758	7.3	1,981	8.8	10,438	8.7	10,683	5.2	15,884	4.2

(Continued on following page)

TABLE 1. Demographic, Comorbid, and Treatment Characteristics of 766,303 Individuals From SEER-Medicare With Selected Cancers and Without Cancer (Continued)

	Bladder Cancer		Other Urinary Cancers[a]		Abdominal Cancers[a]		Other Cancers[a]		Cancer-Free Controls	
	n = 37,940		n = 22,389		n = 120,450		n = 206,125		n = 379,399	
Patient Characteristics	**No.**	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**
Ever smoking	17,176	45.3	7,962	35.6	34,635	28.8	57,918	28.1	62,178	16.4
Moderate-severe liver disease	57	0.2	66	0.3	1,639	1.4	219	0.1	524	0.1
Mild liver disease	194	0.5	162	0.7	4,262	3.5	913	0.4	1,405	0.4
Paralysis (hemiplegia or paraplegia)	242	0.6	167	0.7	982	0.8	883	0.4	1,983	0.5
Peripheral vascular disease	4,935	13.0	2,548	11.4	13,519	11.2	13,771	6.7	23,012	6.1
Moderate-severe renal disease	4,074	10.7	3,272	14.6	11,423	9.5	11,694	5.7	19,877	5.2
Rheumatologic disease	792	2.1	586	2.6	2,876	2.4	3,714	1.8	5,984	1.6
Peptic ulcer disease	367	1.0	257	1.1	2,606	2.2	1,230	0.6	2,318	0.6
Cancer treatment within 1 year of cancer diagnosis										
Cystectomy	3,278	8.6	77	0.3	36	0.0	44	0.0	[c]	0.0
Partial cystectomy	496	1.3	221	1.0	240	0.2	54	0.0	[c]	0.0
BCG	10,497	27.7	377	1.7	36	0.0	141	0.1	34	0.0
TURBT	33,972	89.5	1,812	8.1	626	0.5	2,531	1.2	520	0.1
Carboplatin	1,716	4.5	534	2.4	1,429	1.2	3,146	1.5	101	0.0
Cisplatin	1,795	4.7	291	1.3	2,151	1.8	229	0.1	32	0.0
Thiotepa	234	0.6	[c]	0.0	[c]	0.0	[c]	0.0	0	0.0
Doxorubicin	354	0.9	84	0.4	1,175	1.0	5,577	2.7	35	0.0
Etoposide	196	0.5	21	0.1	303	0.3	183	0.1	19	0.0
Fluorouracil	118	0.3	34	0.2	17,888	14.9	1,505	0.7	57	0.0
Gemcitabine	2,476	6.5	746	3.3	9,561	7.9	475	0.2	68	0.0
Methotrexate	158	0.4	40	0.2	53	0.0	991	0.5	77	0.0
Mitomycin	5,341	14.1	242	1.1	489	0.4	156	0.1	91	0.0
Vinblastine	138	0.4	38	0.2	[c]	0.0	17	0.0	[c]	0.0
Radiation	608	1.6	480	2.1	1,969	1.6	3,830	1.9	224	0.1

Abbreviations: API, Asian Pacific Islander; BCG, Bacille Calmette-Guerin; COPD, chronic obstructive pulmonary disease; MI, myocardial infarction; TURBT, transurethral resection of bladder tumor.

[a] Other urinary cancers are kidney, ureter, and other urinary organs; abdominal cancers are stomach, liver, pancreas, colon, rectum, and gallbladder; other cancers are prostate and female breast.

[b] Summary stage 2000 is not available for gallbladder, other urinary organs, or ureter cancers.

[c] Data suppressed for confidentiality.

and had continuous Medicare parts A and B enrollment in the year following the index date.

For the purpose of estimating potential missed cases of bladder cancer, we chose the classifier with the highest positive predictive value (PPV) and applied it to the noncancer sample. In this context, the PPV quantifies the percentage of correctly identified bladder cancers among those predicted to have bladder cancer by our model. Therefore, we based our estimation of potentially missed bladder cancer cases on the classifier with the highest PPV to provide the most conservative estimate of the number of missed bladder cancer cases. The denominator from the noncancer sample was obtained using the same exclusion criteria as the cases. Starting from 2007 through 2013, we identified the annual number of people included in the noncancer sample who were 65+ years of age and enrolled in Medicare at the first of the year.

The incidence of missed bladder cancer cases was compared with the crude annual incidence rate of bladder cancer among those 65+ years of age reported to SEER. The crude incidence rate was estimated using SEER*Stat.[25] Finally, since we were using the 5% sample, the number of missed cases was multiplied by 20 to scale to the SEER-Medicare noncancer population and used to estimate the total and percentage of bladder cancer cases that were unreported to the registries.

RESULTS

A total of 37,940 patients with bladder cancer were included (Table 1). Most patients were diagnosed with noninvasive (50%) or localized bladder cancer (36%). Compared to the control groups, patients with bladder cancer were more likely to be male and older. The noncancer group was less likely to have comorbid conditions and less likely to have been smokers compared with the bladder cases. Patients with bladder cancer were more likely to have undergone treatment with BCG, cystectomy, TURBT, and mitomycin compared with those diagnosed with other cancers or those without cancer. For example, almost 90% of bladder cancer cases received TURBT treatment within 1 year of diagnosis compared with 8% of those diagnosed with other urinary cancers. Approximately 9% of bladder cancer cases received a cystectomy within a year of diagnosis compared with < 1% of other cancer cases or cancer-free controls.

All five machine learning methods classified bladder cancer cases with very high accuracy (Table 2). Distinguishing bladder cancer from other urinary cancers was the most difficult with an overall accuracy of about 70% across all methods. The overall accuracy for distinguishing bladder from abdominal cancer was ≥ 86%, and more than 91% for bladder cancer versus other unrelated cancers. The highest accuracy was achieved for distinguishing bladder cancer from cancer-free controls, at 95% or higher across all methods.

Sensitivity and specificity were also fairly high across all methods and comparison groups (Table 2). The sensitivity for distinguishing bladder cancer versus other urinary cancers was high (range, 77.8%-81.2%). The sensitivity for distinguishing bladder cancer versus other control groups was higher than that for other urinary cancers. Specificity, which is the probability of correctly predicting that a patient was not diagnosed with bladder cancer, was very high, more than 99%, across all methods and for all comparison groups except other urinary cancers. Specificity was the lowest for classifying bladder cancer versus other urinary cancers (range, 61.5%-72.0%). There was not much variation in the F1 statistic across methods, although the random forest tended to give slightly higher values. The area under the curve was highest for logistic regression compared with the other methods.

CART, random forest, and logistic regression identified BCG, mitomycin, and cystectomy as the most important variables to distinguish bladder cancer versus any of the comparison groups. Figure 1 shows the final classification tree used to classify bladder cancer versus noncancer controls and Figure 2 is the variable importance plot from the random forest. Based on variable importance from random forest, receipt of BCG was the most important variable, followed by mitomycin and cystectomy, which had substantially lower importance compared with BCG. BCG had the largest odds ratio followed by cystectomy in the logistic regression. The SVM assigned largest weights to the use of certain chemotherapies (ie, cisplatin, carboplatin, gemcitabine, and doxorubicin) followed by cystectomy and BCG.

For all comparison groups, logic regression identified interactions between ever having a cystectomy, BCG, and mitomycin. The classification rule for distinguishing bladder cancer from other abdominal cancers was having at least one of cystectomy, BCG, or mitomycin within a year of diagnosis. The classification rules for bladder cancer versus other groups were more complex. For example, the classification rule for identifying bladder cancer versus other urinary cancers included age, sex, and receipt of cisplatin, whereas that for distinguishing bladder cancer from other unrelated cancers included gemcitabine. Finally, if a patient had one of cystectomy, mitomycin, cisplatin, BCG, or carboplatin, then they were classified as having bladder cancer versus a cancer-free control (Fig 3). From 2007 to 2013, there were 165 missed bladder cancer cases identified in the 5% noncancer control sample. The classification tree was used to estimate the potential number of missed cases since it had the highest PPV. The annual rate of missed incident bladder cancer ranged from 16.5 per 100,000 in 2007 to 9.6 per 100,000 in 2013. In comparison, there were 90,714 bladder cancer cases reported to the SEER registries among persons 65+ years of age during that time. This resulted in crude incidence rates ranging from 128.4 per 100,000 in 2007 to 118.1 per 100,000 in 2013. Inclusion of the missed bladder cancer cases would increase the cases reported to SEER by 3.5% from 2007 to 2013. The increase was 6.2% in 2007 declining to 2.6% in 2013.

DISCUSSION

Bladder cancer cases were identified with high accuracy using our machine learning–based rules. BCG is a unique therapeutic procedure for bladder cancer and emerged as the most important variable to distinguish patients with bladder cancer versus other urologic, abdominal, and unrelated cancers followed by mitomycin and cystectomy. Indeed, using CART, we found that approximately 3,300 potential bladder cancer cases were unreported to SEER from 2007 to 2013. Overall, these cases would have increased the SEER reported incidence rate by 3.5%. Registries that have access to claims or other treatment data for cancer cases and the general population may use this classifier to flag cases and follow-up with outpatient facilities. Moreover, some registries may be able to implement case tracking with claims processed in real time. These results are in accordance with standards used for clinical treatment of bladder cancer, namely, BCG and cystectomy being primary treatments.[26]

All machine learning methods had very high sensitivity and specificity in distinguishing bladder cancer cases from the cancer-free controls. CART and logic regression are

TABLE 2. Overall Accuracy, Sensitivity, Specificity, AUC, and F1 Score of Each Classifier

Method	Overall Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	F1
Bladder cancer *v* other urinary cancer					
Classification tree	69.4	79.4	69.1	0.753	0.742
Random forest	72.4	77.8	61.5	0.779	0.783
Logic tree	71.1	81.2	72.0	0.765	0.755
SVM	71.9	79.5	67.1	0.786	0.770
Logistic regression	72.3	78.0	62.3	0.804	0.781
Bladder cancer *v* abdominal cancer					
Classification tree	86.2	96.7	99.5	0.716	0.603
Random forest	87.0	96.8	99.5	0.761	0.633
Logic tree	86.2	96.7	99.5	0.717	0.603
SVM	86.9	96.7	99.5	0.792	0.630
Logistic regression	86.5	93.5	99.0	0.861	0.622
Bladder cancer *v* other cancers					
Classification tree	91.2	98.1	99.8	0.731	0.606
Random forest	91.8	95.9	99.6	0.786	0.649
Logic tree	91.4	95.6	99.6	0.733	0.629
SVM	91.7	94.9	99.5	0.768	0.648
Logistic regression	91.8	95.0	99.5	0.859	0.649
Bladder cancer *v* cancer-free controls					
Classification tree	94.9	99.1	100.0	0.732	0.608
Random forest	95.4	97.7	99.9	0.778	0.666
Logic tree	95.3	98.6	99.9	0.742	0.649
SVM	95.4	97.4	99.9	0.769	0.666
Logistic regression	95.4	97.7	99.9	0.877	0.666

Abbreviations: AUC, area under the curve; SVM, support vector machine.

preferred because of their tree-based structures (lending to easy interpretation), as well as ease of implementation. Therefore, these methods are more amenable to real-time implementation.

Medical claims have been used extensively to identify treatments and cancer diagnoses and our results are concordant with prior studies. One study used automated software for processing billing data from community urology practices to identify an additional 12% of bladder cancer cases that were unreported to the central registry.[5] Lam et al[26] recently developed an algorithm to identify missed cases of bladder cancer using SEER-Medicare data. The algorithm, based on clinical expertise, uses combinations of diagnosis codes, treatment, procedures, and oncology consultations to identify bladder cancer cases. They found about 4% of cases were missed in SEER from 2008 to 2015.

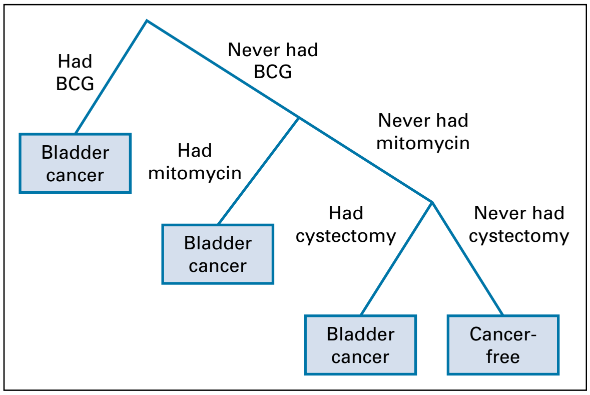

FIG 1. Classification tree to identify bladder cancer cases versus cancer-free controls. BCG, Bacille Calmette-Guerin.

Including TURBT increased predictive performance across all models and all metrics, compared with not including TURBT. However, TURBT is used as a confirmatory procedure for bladder cancer. The prevalence of TURBT among bladder cancer cases is extremely high (90%). So, using TURBT to predict bladder cancer would, by definition, result in a highly sensitive tool. Indeed, secondary analyses including TURBT resulted in TURBT being the **only** variable to distinguish bladder cancers versus other cancers with high sensitivity. However, as mentioned, TURBT is used for diagnosis when a patient is suspected to have bladder cancer. Many patients undergoing the

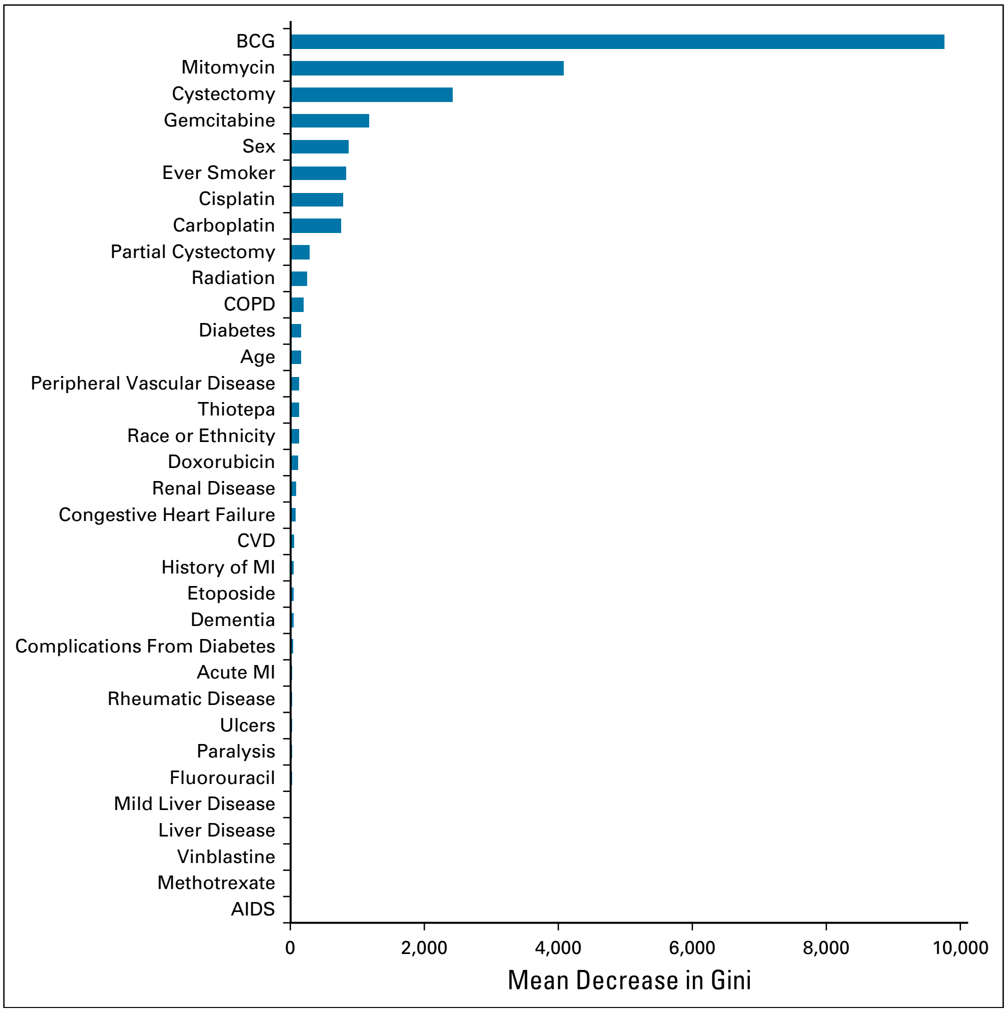

FIG 2. Variable importance plot from the random forest to identify bladder cancer cases versus cancer-free controls. BCG, Bacille Calmette-Guerin; COPD, chronic obstructive pulmonary disease; CVD, cardiovascular disease; MI, myocardial infarction.

procedure may not have bladder cancer. Therefore, using only TURBT to predict bladder cancer would not be useful clinically, since patients who underwent TURBT for suspicion of bladder cancer but did not have it confirmed on TURBT would result in false positives. For this reason, we decided to exclude TURBT from the analyses.

Strengths of our study include a large data set with many cancer cases along with a comprehensive list of treatment and comorbid conditions. We used several comparison groups to classify bladder cancer to challenge the algorithm to distinguish bladder cancer versus other cancers that may have similar profiles. Limitations

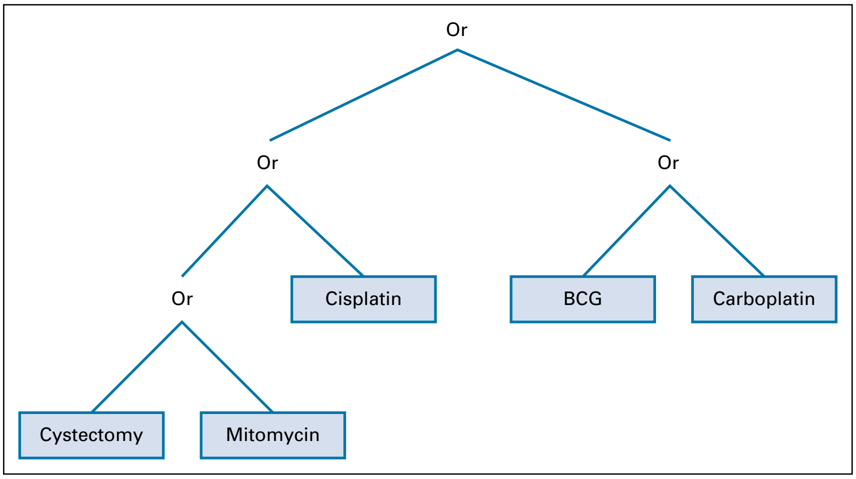

FIG 3. Logic tree to identify bladder cancer cases versus cancer-free controls. If a patient had one of cystectomy, mitomycin, cisplatin, BCG, or carboplatin, then they were classified as having bladder cancer versus being a cancer-free control. BCG, Bacille Calmette-Guerin.

include age restriction (≥ 65 years) for Medicare eligibility. However, the majority of bladder cancer cases occur after 65 years of age. Also, patients with HMO insurance and multiple cancers were excluded, which may limit generalizability. Some treatments or comorbid conditions may have been missed if a claim for payment was not processed through Medicare. Finally, some cancers such as upper tract urothelial carcinoma have similar treatment modalities as bladder cancer. This may have affected our prediction performance when distinguishing bladder cancer from other urologic cancers.

In summary, using machine learning methods, we identified common treatments as the most important variables in distinguishing individuals with bladder cancer compared to those with other cancers or without cancer with very high accuracy. Our results validate what is known clinically about the treatment of bladder cancer and therefore may be useful to cancer registries in identifying cases that may have been unreported to the cancer registry.

AFFILIATIONS

[1]Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD
[2]University of North Carolina Lineberger Comprehensive Cancer Center, Chapel Hill, NC
[3]Biostatistics and Clinical Data Management Core, University of North Carolina Lineberger Comprehensive Cancer Center, Chapel Hill, NC
[4]Information Management Services Inc, Calverton, MD
[5]University of Michigan, Ann Arbor, MI

CORRESPONDING AUTHOR

Anne-Michelle Noone, MS, Data Analytics Branch, Surveillance Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, 9609 Medical Center Dr, 4E-552, Rockville, MD 20850; e-mail: nonea@mail.nih.gov.

AUTHOR CONTRIBUTIONS

Conception and design: Anne-Michelle Noone, Clara J. K. Lam, Angela B. Smith, Angela B. Mariotto, Mousumi Banerjee
Collection and assembly of data: Anne-Michelle Noone, Eric Boyd
Data analysis and interpretation: Anne-Michelle Noone, Clara J. K. Lam, Angela B. Smith, Matthew E. Nielsen, Mousumi Banerjee
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by the authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Angela B. Smith
Consulting or Advisory Role: Merck, Photocure, Urogen Pharma, Fergene, Ambu

Matthew E. Nielsen
Stock and Other Ownership Interests: Grand Rounds Health

No other potential conflicts of interest were reported.

REFERENCES

1. Sherman R, Firth R, De P, et al (eds): Cancer in North America: 2012-2016, in Combined Cancer Incidence for the United States, Canada and North America, Volume 1. Springfield, IL, North American Association of Central Cancer Registries, 2019
2. Parkin DM, Bray F: Evaluation of data quality in the cancer registry: Principles and methods Part II. Completeness. Eur J Cancer 45:756-764, 2009
3. Hofferkamp J (ed): Standards for Cancer Registries Volume III: Standards for Completeness, Quality, Analysis, Management, Security and Confidentiality of Data. Springfield, IL, North American Association of Central Cancer Registries, 2008
4. Das B, Clegg LX, Feuer EJ, et al: A new method to evaluate the completeness of case ascertainment by a cancer registry. Cancer Causes Control 19:515-525, 2008
5. Penberthy LT, McClish D, Agovino P: Impact of automated data collection from urology offices: Improving incidence and treatment reporting in urologic cancers. J Registry Manag 37:141-147, 2010
6. Snyder C, Harlan L, Knopf K, et al: Patterns of care for the treatment of bladder cancer. J Urol 169:1697-1701, 2003
7. Babjuk M, Bohle A, Burger M, et al: EAU guidelines on non-muscle-invasive urothelial carcinoma of the bladder: Update 2016. Eur Urol 71:447-461, 2017
8. American Cancer Society: Treating Bladder Cancer. https://www.cancer.org/cancer/bladder-cancer/treating.html
9. National Cancer Institute: SEER-Medicare Linked Database. https://healthcaredelivery.cancer.gov/seermedicare/
10. Surveillance, Epidemiology and End Results Program: Incidence Site Recode Variables. National Cancer Institute. http://seer.cancer.gov/siterecode/
11. Breiman L, Friedman JH, Olshen RA, et al: Classification and Regression Trees. Belmont, CA, Wadsworth, 1984
12. Banerjee M, Reynolds E, Andersson HB, et al: Tree-based analysis: A practical appraoch to create clinical decision-making tools. Circ Cardiovasc Qual Outcomes 12:e004879, 2019
13. Noone AM, Banerjee M: Machine learning methods for cancer diagnosis and prognostication, in Khattree R, Nail D (eds): Computational Methods in Biomedical Research. London, United Kingdom, Chapman & Hall, 2008
14. Breiman L: Random forests. Machine Learn 45:5-32, 2001
15. Ruczinski I, Kooperberg C, Leblanc M: Logic regression. J Comput Graphical Stat 12:475-511, 2003
16. Banerjee M, Filson C, Xia R, et al: Logic regression for provider effects on kidney cancer treatment delivery. Comput Math Methods Med 2014:316935, 2014

17. Reynolds E, Callaghan B, Banerjee M: SVM-CART for disease classification. J Appl Stat 46:2987-3007, 2019
18. Hastie T, Tibshirani R, Friedman JH: The Elements of Statistical Learning : Data Mining, Inference, and Prediction (ed 2). New York, NY, Springer, 2009
19. Sokolova M, Japkowicz N, Szpakowicz S: Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. Berlin, Heidelberg, Springer, 2006
20. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, 2018. https://www.R-project.org/
21. rpart: Recursive Partitioning and Regression Trees, 2018. https://CRAN.R-project.org/package=rpart
22. Classification and Regression by randomForest. 2002. https://CRAN.R-project.org/doc/Rnews/
23. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. 2018. https://CRAN.R-project.org/package=e1071
24. LogicReg: Logic Regression, 2019. https://CRAN.R-project.org/package=LogicReg
25. Surveillance Research Program: NCISSswscgsv
26. Lam C, Warren JL, Neilsen ME, et al: Using the SEER-Medicare data to assess incident chronic myeloid leukemia and bladder cancer cases missed by Cancer Registries. J Natl Cancer Inst Monogr 2020:31-38, 2020

APPENDIX

TABLE A1. List of CPT and ICD-9 Codes to Identify Treatment

Treatment	CPT Codes	ICD-9 Codes
Cystectomy	51570, 51575, 51580, 51585, 51590, 51595, 51596, 51597	57.71, 57.7
TURBT or bladder biopsy	52234, 52235, 52240, 52204, 52214, 52224, 52305	57.33, 57.34
Partial cystectomy	51550, 51555, 51565	57.6
Chemotherapy	51720	99.25, V58.1, V58.22, V66.2, V67.2
Radiation	77261, 77399, 77400, 77490, 77750, or 77797	92.1-92.9, V58.0, V66.1, V67.1

Abbreviations: CPT, Current Procedures Terminology; ICD, International Classification of Diseases; TURBT, transurethral resection of bladder tumor.

TABLE A2. List of HCPC to Identify Chemotherapy

HCPC	Generic Name	Description
J9045	Carboplatin	Injection, carboplatin, 50 mg
C9418	Cisplatin	Cisplatin, powder or solution, per 10 mg, brand name
J9060	Cisplatin	Cisplatin, powder or solution, per 10 mg
J9062	Cisplatin	Cisplatin, 50 mg
C9415	Doxorubicin HCl	Doxorubicin HCl, 10 mg, brand name
J9000	Doxorubicin HCl	Injection, doxorubicin HCl, 10 mg
Q2048	Doxorubicin HCl	Injection, doxorubicin HCl, 10 mg
J9001	Doxorubicin HCl liposomal	Injection, doxorubicin HCl, all lipid formulations, 10 mg
Q2049	Doxorubicin HCl liposomal	Injection, doxorubicin HCl, all lipid formulations, 10 mg
C9414	Etoposide	Etoposide; oral, 50 mg, brand name
C9425	Etoposide	Etoposide, 10 mg, brand name
J8560	Etoposide	Etoposide; oral, 50 mg
J9181	Etoposide	Injection, etoposide, 10 mg
J9182	Etoposide	Etoposide, 100 mg
J9190	Fluorouracil	Injection, fluorouracil, 500 mg
J9201	Gemcitabine	Injection, gemcitabine HCl, 200 mg
J8610	Methotrexate	Methotrexate; oral, 2.5 mg
J9250	Methotrexate	Methotrexate sodium, 5 mg
J9260	Methotrexate	Methotrexate sodium, 50 mg
C9432	Mitomycin	Mitomycin, 5 mg, brand name
J9280	Mitomycin	Mitomycin, 5 mg
J9290	Mitomycin	Mitomycin, 20 mg
J9291	Mitomycin	Mitomycin, 40 mg
J9360	Vinblastine sulphate	Injection, vinblastine sulfate, 1 mg
J9031	BCG	BCG immunotherapy
J9340	Thiotepa	Injection, thiotepa, 15 mg

Abbreviations: BCG, Bacille Calmette-Guerin; HCPC, Healthcare Common Procedure Code.

TABLE A3. List of Parameters and R Function Used for Each Method

Method	Variables	R Function	Tuning Parameters
Classification tree	All models used the variables in Table 1 except TURBT	cart	Min No. of observations in a node to attempt split (minsplit): 20 Min No. of observations in terminal node: minsplit/3 No. of cross-validations: 10
Random forest	All models used the variables in Table 1 except TURBT	randomForest	No. of variables randomly sampled as candidates at each split: sqrt(number of covariates) No. of trees: 500
Support vector machine	All models used the variables in Table 1 except TURBT	e1071	Kernel: radial
Logic regression	All models used the variables in Table 1 except TURBT	logicReg	No. of trees: 1 Model selection: cross-validation No. of groups cases are assigned to: 15 Maximum No. of leaves to be fit in all trees combined: 15

Abbreviation: TURBT, transurethral resection of bladder tumor.