Cancer Epidemiology 61 (2019) 30–37

Contents lists available at ScienceDirect

Cancer Epidemiology

journal homepage: www.elsevier.com/locate/canep

Development of predictive models to identify advanced-stage cancer patients in a US healthcare claims database

Daina B. Esposito[a,b,1], Leo Russo[c], Dina Oksen[d,1], Ruihua Yin[a], Vibha C.A. Desai[a,1], Jennifer G. Lyons[a,1], Patrice Verpillat[d], Jose L. Peñalvo[d,1], Francois-Xavier Lamy[d], Stephan Lanes[a,*]

[a] HealthCore, Inc., Wilmington, DE, United States
[b] Boston University, Boston, MA, United States
[c] Pfizer, Inc., Collegeville, PA, United States
[d] Merck KGaA, Darmstadt, Germany

ARTICLE INFO

Keywords:
Ovarian cancer
Urothelial carcinoma
Gastric adenocarcinoma
Merkel cell carcinoma
Non-small cell lung cancer
Predictive modeling
Machine learning
LASSO regression

ABSTRACT

Background: Although healthcare databases are a valuable source for real-world oncology data, cancer stage is often lacking. We developed predictive models using claims data to identify metastatic/advanced-stage patients with ovarian cancer, urothelial carcinoma, gastric adenocarcinoma, Merkel cell carcinoma (MCC), and non-small cell lung cancer (NSCLC).

Methods: Patients with ≥1 diagnosis of a cancer of interest were identified in the HealthCore Integrated Research Database (HIRD), a United States (US) healthcare database (2010–2016). Data were linked to three US state cancer registries and the HealthCore Integrated Research Environment Oncology database to identify cancer stage. Predictive models were constructed to estimate the probability of metastatic/advanced stage. Predictors available in the HIRD were identified and coefficients estimated by Least Absolute Shrinkage and Selection Operator (LASSO) regression with cross-validation to control overfitting. Classification error rates and receiver operating characteristic curves were used to select probability thresholds for classifying patients as cases of metastatic/advanced cancer.

Results: We used 2723 ovarian cancer, 6522 urothelial carcinoma, 1441 gastric adenocarcinoma, 109 MCC, and 12,373 NSCLC cases of early and metastatic/advanced cancer to develop predictive models. All models had high discrimination (C > 0.85). At thresholds selected for each model, PPVs were all > 0.75: ovarian cancer = 0.95 (95% confidence interval [95% CI]: 0.94–0.96), urothelial carcinoma = 0.78 (95% CI: 0.70–0.86), gastric adenocarcinoma = 0.86 (95% CI: 0.83–0.88), MCC = 0.77 (95% CI 0.68–0.89), and NSCLC = 0.91 (95% CI 0.90 – 0.92).

Conclusion: Predictive modeling was used to identify five types of metastatic/advanced cancer in a healthcare claims database with greater accuracy than previous methods.

1. Background

Health insurance claims databases are a valuable source of real world data in oncology. With systematic and longitudinal capture of data for millions of patients, they offer opportunities to better characterize the natural history of various cancer types, and assess rare outcomes. Data are collected on parameters across healthcare providers and settings, including diagnoses, procedures, medications, site of care, and provider specialty. However, healthcare claims databases lack information on cancer stage, which constitutes a considerable limitation in identifying a target population of a treatment intended for patients with metastatic or advanced stage cancer.

Past efforts to identify cancer stage in healthcare claims databases have met limited success [1–5]. For example, a recent validation study assessing the feasibility of identifying cancer progression to metastatic stage based on algorithms that selected variables relying on both clinician input and random forest methods obtained positive predictive values (PPVs) below 50% for multiple sites [1]. Similarly, low PPVs

* Corresponding author at: Wilmington Office, 123 Justison Street, Suite 200, Wilmington, DE 19801, United States.
E-mail address: slanes@healthcore.com (S. Lanes).
[1] At time of study.

https://doi.org/10.1016/j.canep.2019.05.006
Received 20 September 2018; Received in revised form 21 March 2019; Accepted 10 May 2019
Available online 22 May 2019
1877-7821/

were reported for algorithms that depend on secondary malignancy codes [2]. A study of linked Surveillance, Epidemiology and End Results (SEER)-Medicare data also had low PPVs and poor sensitivity for algorithms identifying advanced stage cancer [5]. Better algorithm performance has been observed with cancer site-specific versus general algorithms [4,6], and model-based algorithms have shown that it is feasible to identify cancer recurrence [7]. Regardless, several authors have concluded that advanced cancer stage cannot be inferred in healthcare claims databases with reasonable accuracy. As an alternative, confirmation of advanced stage for all patients included in a study using additional clinical data sources such as registries and medical records has been proposed [1]. Many healthcare claims databases are de-identified, however, and are not able to be linked to external sources containing clinical data for validation. For identifiable claims databases, the large populations mean this strategy would add considerable and sometimes prohibitive cost and time to study execution. Predictive modeling uses clinical data as a validation sample to provide reliable case classification as an outcome in constructing an algorithm that can then be applied to identify cases in the large source population for whom clinical data are unavailable.

In this study, we used predictive modeling to develop validated algorithms that identify five types of advanced or metastatic or advanced stage cancer (ovarian cancer, urothelial carcinoma, gastric adenocarcinoma, Merkel cell carcinoma [MCC], and non-small cell lung cancer [NSCLC]). These cancer sites were selected in anticipation of the need for real-world data characterizing the target populations for a monoclonal antibody indicated for advanced stage cancer.

2. Methods

This study was conducted using the HealthCore Integrated Research Database (HIRD), a large healthcare claims database in the US. The HIRD contains professional claims, facility claims, outpatient pharmacy claims, outpatient laboratory results, and enrollment information obtained from affiliated health plans across the country. Personally identifying information is captured in the database, and each individual can be tracked across all healthcare encounters for which claims are processed throughout the course of their health plan enrollment. The HIRD includes enrollment data, medical care, prescription drug use, and healthcare utilization. Diagnoses and procedures received during physician, specialist, and emergency room visits, as well as hospital stays, are captured in the database using International Classification of Diseases, tenth edition Clinical Modification (ICD-10-CM) for diagnosis coding and Procedure Coding System (ICD-10-PCS) for procedure coding (since 01 October 2015) and ninth edition (ICD-9-CM) diagnostic, ICD-9-PCS procedure, Current Procedural Terminology (CPT) procedures, HealthCare Common Procedure Coding System (HCPCS), and place of service codes. Outpatient pharmacy claims are captured by National Drug Codes (NDCs). Information on physician specialty and laboratory result data are also retained in the database.

Patients at least 18 years of age with at least one ICD-9-CM or ICD-10-CM diagnosis code for one of the five cancers of interest between 2010 and 2016 and at least six months of health plan enrollment were matched to two sources of data to identify cancer stage: (1) cancer registry data for three participating states (Ohio, Kentucky, and New York), and (2) HealthCore Integrated Research Environment Oncology (HIRE-O) data. Patients were excluded from the validation study if they were not successfully matched to a participating cancer registry in the state of their residence, if there was disagreement between the primary cancer site identified in the HIRD compared to that identified in either the HIRE-O database or the cancer registry, or if there was inadequate staging information to determine if the patient was an metastatic/advanced or early case.

In the participating cancer registries, primary cancer site for ovarian cancer and gastric adenocarcinoma was determined using International Classification of Disease for Oncology, Third Edition (ICD-O-3) codes for primary site location. To identify primary site for urothelial carcinoma, MCC, and NSCLC, ICD-O-3 histology codes were used in addition to the ICD-O-3 site codes. Given the cross-sectional nature of registry data, stage was captured at initial diagnosis.

The HIRE-O is a novel program designed to align reimbursement with evidence-based, cost-effective oncology treatment. Clinical data obtained from the HIRE-O program includes cancer type, pathology, stage, and selected biomarker data. The data are recorded by the physician and stored electronically so they can be linked on a patient-level with data from the HIRD. The HIRE-O program enables physicians to obtain additional reimbursement for following preferred treatment regimens that encourage evidence-based quality care for the patients. Clinical data were obtained when physicians participated in the HIRE-O reimbursement program or sought prior authorization for various cancer treatments. Results of validation studies comparing clinical information in HIRE-O against medical record review suggested good agreement between sources for a variety of metrics [8]. Where multiple records of stage were available, the date of the first diagnosis of metastatic/advanced cancer stage was used in analyses.

Metastatic or advanced cancer was defined as: epithelial stage III and IV disease for ovarian cancer; locally advanced (T4b, N2 or N3) or metastatic (M1) disease for urothelial carcinoma; unresectable, locally advanced or metastatic disease for gastric adenocarcinoma; locally advanced (stage III) or metastatic (stage IV) disease for MCC; and locally advanced (stage IIIa or IIIb) or metastatic (stage IV) disease for NSCLC. In the registry data, cancer stage was assigned as metastatic/advanced or early in a stepwise fashion to account for possible missing data in the cancer registries. First, American Joint Committee on Cancer (AJCC) cancer staging was used to define stage. If those variables were missing, the next step assigned stage based on the reported clinical or pathology reported stage group. For urothelial carcinoma, the next step utilized derived AJCC tumor-node-metastasis (TNM) cancer staging variables to assign stage based on T, N, or M stage individually. Next, SEER summary stage was used to define stage categories as the SEER summary stage variable is less precise than those used in earlier steps [9]. In the final step, ICD-O-3 behavior codes were used to define metastatic/advanced stage. If a patient could not be assigned to either early or metastatic/advanced stage cancer, they were assigned as unknown cancer stage (Supplemental Table 1).

After obtaining cancer staging from the cancer registries and HIRE-O, predictive model algorithms were developed and validated. For each cancer type, the predictive model used various characteristics in the claims to estimate the probability of being a case of metastatic/advanced cancer. We utilized least absolute shrinkage and selection operator (LASSO) regression analysis [10] to perform model variable selection and cross-validation. The LASSO method reduces overfitting by constraining the sum of regression coefficients in selecting covariates for inclusion in the model and estimating coefficients for each covariate. This method addresses collinearity when evaluating a large number of potential covariates relative to the number of observations [10]. Within each metastatic/advanced cancer population, potential covariates included age, gender, frequently occurring diagnoses and medications, radiation history, and surgeries performed at times surrounding the cancer diagnosis. The predictors identified and included in the algorithms were empirically defined based on their prediction of metastatic/advanced cancer stage, and do not suggest causality.

Each model with LASSO regression was tested by cross-validation using the function glmnet [11] in R that performs K-fold cross validation to account for optimistic estimation [10] with K of 20. In 20-fold cross-validation, the sample was divided into 20 subsets. For each of the 20 training iterations, each of the subsets was excluded once for use in validation, and the model was fit with the remaining 19 subsets. The results from the 20 training iterations were used to select the shrinkage operator and the best-fitting model to minimize the prediction error rate. Predictive model coefficients were presented and performance characteristics were described, and charted, across cancer probability

scores. Probability thresholds were selected based on performance characteristics with the goal of identifying a balance between defining a study entry criterion that is adequately inclusive for a future cohort of metastatic/advanced cancer patients, and the competing goal of minimizing false positives. Receiver operator characteristic (ROC) curves were created to evaluate model performance. In the final step, we applied the models to patients in the database with at least six months of continuous enrollment and at least one diagnosis code for the cancer of interest to identify for each site a large population of patients with metastatic/advanced cancer.

This study was approved by the New England Institutional Review Board, and all analyses were performed using R Version 3.2.5 and SAS Enterprise Guide 5.1 (SAS Institute, Cary, NC).

3. Results

There were 191,313 patients identified in the HIRD with a cancer diagnosis of interest from 01 January 2010 through 30 September 2016. Of these, 28,458 had a diagnosis of ovarian cancer, 58,166 had a diagnosis of urothelial cell carcinoma, 14,461 had a diagnosis of gastric adenocarcinoma, 1114 had a diagnosis of MCC, and 89,114 had a diagnosis of NSCLC. After matching to a registry or HIRE-O, applying inclusion and exclusion criteria, and limiting to patients with confirmed stage data, the validation sample included 1978 metastatic/advanced ovarian cancer patients, 791 metastatic/advanced urothelial carcinoma patients, 937 metastatic/advanced gastric adenocarcinoma patients, 48 metastatic/advanced MCC patients, and 8516 metastatic/advanced NSCLC patients (Fig. 1).

3.1. Ovarian Cancer

Metastatic/advanced ovarian cancer patients (n = 1978) had a mean age of 59.7 years (standard deviation 12.09, Table 1).

The LASSO regression model performed well across all probability thresholds, with a cross-validated PPV ranging from 0.73 to 0.97 (Table 2). The coefficients from the LASSO regression model can be found in Supplemental Table 2 and are shown on the log-odds scale. Of the 77 variables assigned a non-zero weight by LASSO regression, the strongest predictors of confirmed metastatic/advanced ovarian cancer included a diagnosis of secondary malignant neoplasm of the respiratory and digestive systems (β = 1.47), treatment with bevacizumab (β = 1.01), and treatment with olaparib (β = 0.67). The strongest predictors of confirmed non- metastatic/advanced ovarian cancer included treatment with the "miscellaneous antineoplastics" generic product identifier (GPI) subclass (β = -1.45), treatment with goserelin (β = -0.96), and treatment with chemotherapy rescue/antidote agents (β = -0.74).

The number of early cancer cases incorrectly classified as advanced (i.e., false positives) decreased from 739 at a predicted probability threshold of 0.1–34 at a predicted probability of 0.9 (Fig. 2). Taking into account both Table 2 and Fig. 2, a threshold of 0.80 was selected, with a PPV of 0.95 (95% confidence interval [CI] 0.94 – 0.96) and a sensitivity of 70% (Table 2). This model had a c-statistic of 0.89 indicating a high degree of discrimination between true and false cases of metastatic/advanced ovarian cancer (Fig. 3).

3.2. Urothelial carcinoma

Patients with metastatic/advanced urothelial carcinoma (n = 791) had a mean age of 69.3 years (standard deviation 11.58,) and 30.1% were female (Table 1).

The LASSO regression model performed poorly at lower thresholds and moderately at thresholds above 0.60, with PPVs ranging from 0.12 to 0.94 (Table 2). Of the 84 variables assigned a non-zero weight by LASSO regression, the strongest predictors of confirmed metastatic/advanced urothelial carcinoma included treatment with gemcitabine (β = 0.98), diagnosis of malignant neoplasm of the kidney and other unspecified urinary organs (β = 0.92), and a visit to a hematologist (β = 0.85).The strongest predictors of confirmed non-advanced urothelial carcinoma included treatment with pembrolizumab (β = -1.33), topoisomerase I inhibitors (β = -1.15), and a visit to a urologist (β = -0.43 [Supplemental Table 2]).

There were 1158 false positive cases identified at a predicted probability of metastatic/advanced cancer threshold of 0.10, and this number decreased to < 10 at a threshold of 0.9 (Fig. 2). Taking into account both Table 2 and Fig. 2, a threshold of 0.60 was selected, with a

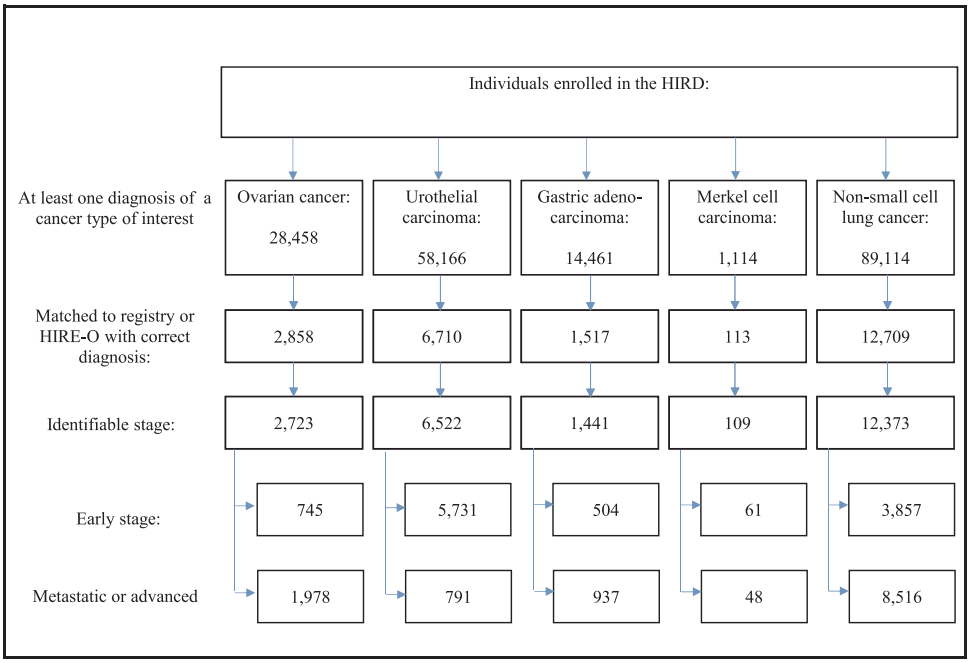

Fig. 1. Study Population.
Abbreviations: HIRD = HealthCore Integrated Research Database, HIRE-O = HealthCore Integrated Research Environment Oncology.

Table 1
Demographic characteristics by cancer site, metastatic/advanced cancer cases.

	Ovarian Cancer		Urothelial Carcinoma		Gastric Adenocarcinoma		Merkel Cell Carcinoma		Non-Small Cell Lung Cancer	
	N	%	N	%	N	%	N	%	N	%
Total cases (N, %)	1,978	72.6	791	29.0	937	65.0	48	44.0	8,516	68.8
Demographic characteristics										
Age (years)										
Mean, standard deviation	59.7	12.09	69.3	11.58	64.9	13.29	71.8	11.64	67.2	11.10
Median	59		70		65		74		67	
Minimum – maximum	11	98	26	96	20	100	44	90	14	99
Age category (N, %)										
Less than 18	< 10		0		0		0		< 10	
18 to 34	37	1.9	< 10		13	1.4	0		26	0.3
35 to 49	605	30.6	73	9.2	189	20.2	< 10		1,104	13.0
50 to 64	752	38.0	199	25.2	253	27.0	11	22.9	2,405	28.2
65 and older	579	29.3	515	65.1	482	51.4	33	68.8	4,979	58.5
Female (N, %)	1,978	100.0	238	30.1	310	33.1	16	33.3	4,052	47.6
US Region of Residence (N, %)										
Northeast	454	23.0	270	34.1	325	34.7	17	35.4	2,638	31.0
Midwest	786	39.7	299	37.8	316	33.7	17	35.4	3,251	38.2
West	395	20.0	167	21.1	187	20.0	10	20.8	2,175	25.5
South	343	17.3	55	7.0	109	11.6	< 10		451	5.3
Unknown	0		0		0		0		< 10	
Year of index date (N, %)										
2010	416	21.0	199	25.2	168	17.9	11	22.9	1,766	20.7
2011	188	9.5	110	13.9	124	13.2	< 10		1,220	14.3
2012	217	11.0	117	14.8	122	13.0	< 10		1,202	14.1
2013	239	12.1	116	14.7	123	13.1	< 10		1,171	13.8
2014	301	15.2	97	12.3	153	16.3	< 10		1,192	14.0
2015	355	18.0	92	11.6	158	16.9	< 10		1,155	13.6
2016	262	13.3	60	7.6	89	9.5	< 10		810	9.5
Duration follow-up (months)										
Mean, standard deviation	22.6	20.32	21.8	20.62	14.7	16.60	25.3	23.01	14.3	16.27
Median	16		15		9		14		8	
Minimum – maximum	0	82	0	82	0	82	3	81	0	82
Deyo-Charlson Comorbidity Index[a]										
Mean, standard deviation	3.2	3.95	3.3	3.66	3.6	3.89	4.8	4.18	3.6	3.76
Median	1		2		2		4		2	
Minimum – maximum	0	15	0	17	0	17	0	15	0	21
Count of distinct drugs used										
Mean, standard deviation	15.1	11.24	14.2	11.14	12.2	9.50	16.0	10.16	13.6	10.97
Median	13		13		11		15		12	
Minimum – maximum	0	74	0	56	0	56	0	40	0	87
At least one surgical procedure (N, %)	1,238	62.6	489	61.8	171	18.2	10	20.8	829	9.7
At least one radiation therapy (N, %)	196	9.9	155	19.6	286	30.5	28	58.3	4,146	48.7
At least one anti-cancer therapy (N, %)	1,630	82.4	482	60.9	626	66.8	24	50.0	5,448	64.0

Abbreviations: N = number; US = United States. Counts below 10 are shown as " < 10″ with percentages omitted to protect patient privacy.

[a] Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J* Clin Epidemiol 1992;45:613–19. doi:10.1016/0895-4356(92)90133-8.

PPV of 0.78 (95% CI 0.73 – 0.83) and a sensitivity of 30% (Table 2). This model had a c-statistic of 0.86 indicating a high degree of discrimination between true and false cases of confirmed metastatic/advanced urothelial carcinoma (Fig. 3).

3.3. Gastric adenocarcinoma

Patients with metastatic/advanced gastric adenocarcinoma (n = 937) had a mean age of 64.9 years (standard deviation 13.29), and 33.1% were females (Table 1).

The LASSO regression model performed moderately across all probability thresholds, with a PPV ranging from 0.65 to 0.97 (Table 2). Of the 66 variables assigned a non-zero weight by LASSO regression, the strongest predictors of confirmed metastatic/advanced gastric adenocarcinoma included treatment with etoposide (β = 1.49), diagnosis of secondary malignant neoplasm of the respiratory and digestive systems (β = 1.14), and treatment with ramucirumab (β = 0.80). The strongest predictors of confirmed non-advanced gastric adenocarcinoma included treatment with everolimus (β = -1.79), treatment with mitomycin (β = -1.02), and endoscopic mucosal resection and endoscopic sub-mucosal dissection (β = -0.97 [Supplemental Table 2]).

The number of false positives decreased from 497 at a predicted probability threshold of 0.1–10 at a predicted probability of 0.9 (Fig. 2). A threshold of 0.60 was selected, with a PPV of 0.86 (95% CI 0.83 – 0.88) and a sensitivity of 79% (Table 2). This model had a c-statistic of 0.86 indicating a high degree of discrimination between true and false cases of confirmed metastatic/advanced gastric adenocarcinoma (Fig. 3).

3.4. Merkel cell carcinoma

Patients with metastatic/advanced MCC (n = 48) had a mean age of 71.8 years (standard deviation 11.64), and 33.3% were females (Table 1).

The LASSO regression model performed poorly at lower thresholds and well at probability thresholds above 0.40, with a PPV ranging from 0.44 to 1.00 (Table 2). Of the 20 variables assigned a non-zero weight by LASSO regression, the strongest predictors of confirmed metastatic/advanced MCC included treatment with alkylating agents (β = 0.74), treatment with somatostatin analogues (β = 0.56), and treatment with antineoplastics (β = 0.51). The strongest predictors of confirmed non-advanced MCC included other and unspecified disorders of the joint (β

Table 2
Cross-validated algorithm performance by site and predicted probability threshold.

	True positives	False positives	True negatives	False negatives	Sensitivity	Specificity	PPV	95% LCL	95% UCL
Predicted probability threshold									
Metastatic/advanced ovarian cancer									
0.0	1978	745	0	0	1.00	0.00	0.73	0.71	0.74
0.1	1978	739	< 10	0	1.00	0.01	0.73	0.71	0.74
0.2	1971	681	64	< 10	1.00	0.09	0.74	0.73	0.76
0.3	1942	534	211	36	0.98	0.28	0.78	0.77	0.80
0.4	1862	367	378	116	0.94	0.51	0.84	0.82	0.85
0.5	1780	233	512	198	0.90	0.69	0.88	0.87	0.90
0.6	1676	172	573	302	0.85	0.77	0.91	0.89	0.92
0.7	1571	116	629	407	0.79	0.84	0.93	0.92	0.94
0.8*	1394	78	667	584	0.70	0.90	0.95	0.94	0.96
0.9	1087	34	711	891	0.55	0.95	0.97	0.96	0.98
1.0	0	0	745	1978	0.00	1.00			
Metastatic/advanced urothelial carcinoma									
0.0	791	5731	0	0	1.00		0.12	0.11	0.13
0.1	653	1158	4573	138	0.83		0.36	0.34	0.38
0.2	543	576	5155	248	0.69		0.49	0.46	0.51
0.3	463	338	5393	328	0.59		0.58	0.54	0.61
0.4	385	203	5528	406	0.49		0.65	0.62	0.69
0.5	299	121	5610	492	0.38		0.71	0.67	0.76
0.6*	241	68	5663	550	0.30		0.78	0.73	0.83
0.7	148	29	5702	643	0.19		0.84	0.78	0.89
0.8	78	< 10	5725	713	0.10		0.93	0.87	0.98
0.9	16	< 10	5730	775	0.02		0.94	0.83	1.05
1.0	0	0	5731	791	0.00				
Metastatic/advanced gastric adenocarcinoma									
0	937	504	0	0	1.00	0.00	0.65	0.63	0.67
0.1	937	497	< 10	0	1.00	0.01	0.65	0.63	0.68
0.2	933	426	78	< 10	1.00	0.15	0.69	0.66	0.71
0.3	911	344	160	26	0.97	0.32	0.73	0.70	0.75
0.4	880	260	244	57	0.94	0.48	0.77	0.75	0.80
0.5	818	182	322	119	0.87	0.64	0.82	0.79	0.84
0.6*	738	125	379	199	0.79	0.75	0.86	0.83	0.88
0.7	639	81	423	298	0.68	0.84	0.89	0.86	0.91
0.8	513	33	471	424	0.55	0.93	0.94	0.92	0.96
0.9	304	10	494	633	0.32	0.98	0.97	0.95	0.99
1.0	0	0	504	937	0.00	1.00			
Metastatic/advanced Merkel cell carcinoma									
0.0	48	61	0	0	1.00	0.00	0.44	0.35	0.53
0.1	48	58	< 10	0	1.00	0.05	0.45	0.36	0.55
0.2	47	44	17	< 10	0.98	0.28	0.52	0.42	0.62
0.3	45	30	31	< 10	0.94	0.51	0.60	0.52	0.75
0.4*	44	13	48	< 10	0.92	0.79	0.77	0.68	0.89
0.5	34	< 10	55	14	0.71	0.90	0.85	0.80	0.99
0.6	29	0	61	19	0.60	1.00	1.00	0.91	1.03
0.7	18	0	61	30	0.38	1.00	1.00	1.00	1.00
0.8	< 10	0	61	41	0.15	1.00	1.00	1.00	1.00
0.9	0	0	61	48	0.00	1.00	1.00	1.00	1.00
1.0	0	0	61	48	0.00	1.00			
Metastatic/advanced non-small cell lung cancer									
0.0	8,516	3,857	0	0	1.00	0.00	0.69	0.68	0.70
0.1	8,476	3,308	549	40	1.00	0.14	0.72	0.71	0.73
0.2	8,339	2,647	1,210	177	0.98	0.31	0.76	0.75	0.77
0.3	8,185	2,100	1,757	331	0.96	0.46	0.80	0.79	0.80
0.4	7,971	1,682	2,175	545	0.94	0.56	0.83	0.82	0.83
0.5	7,703	1,313	2,544	813	0.90	0.66	0.85	0.85	0.86
0.6	7,324	971	2,886	1,192	0.86	0.75	0.88	0.88	0.89
0.7*	6,782	667	3,190	1,734	0.80	0.83	0.91	0.90	0.92
0.8	5,968	393	3,464	2,548	0.70	0.90	0.94	0.93	0.94
0.9	4,415	157	3,700	4,101	0.52	0.96	0.97	0.96	0.97
1.0	0	0	3,857	8,516	0.00	1.00			

Abbreviations: PPV positive predictive value; LCL lower confidence interval; UCL upper confidence interval.
Recommended algorithm thresholds for identification of a cohort of advanced cancer patients are designated with an asterisk (*).

= -0.62), diagnosis of disorders of lipid metabolism (β = -0.54), and other and unspecified malignant neoplasm of skin (β = -0.43 [Supplemental Table 2]).

The number of false positive cases was 58 at a predicted probability threshold of 0.1, and there were no false positives identified above a threshold of 0.60 (Fig. 2). A threshold of 0.40 was selected, with a PPV of 0.77 (95% CI 0.68 – 0.89) and a sensitivity of 92% (Table 2). This model had a c-statistic of 0.94 indicating a high degree of discrimination between true and false cases of confirmed metastatic/advanced MCC (Fig. 3).

3.5. Non-small cell lung cancer

Patients with metastatic/advanced NSCLC (n = 8516) had a mean age of 67.2 years (standard deviation 11.10), and 47.6% were females (Table 1).

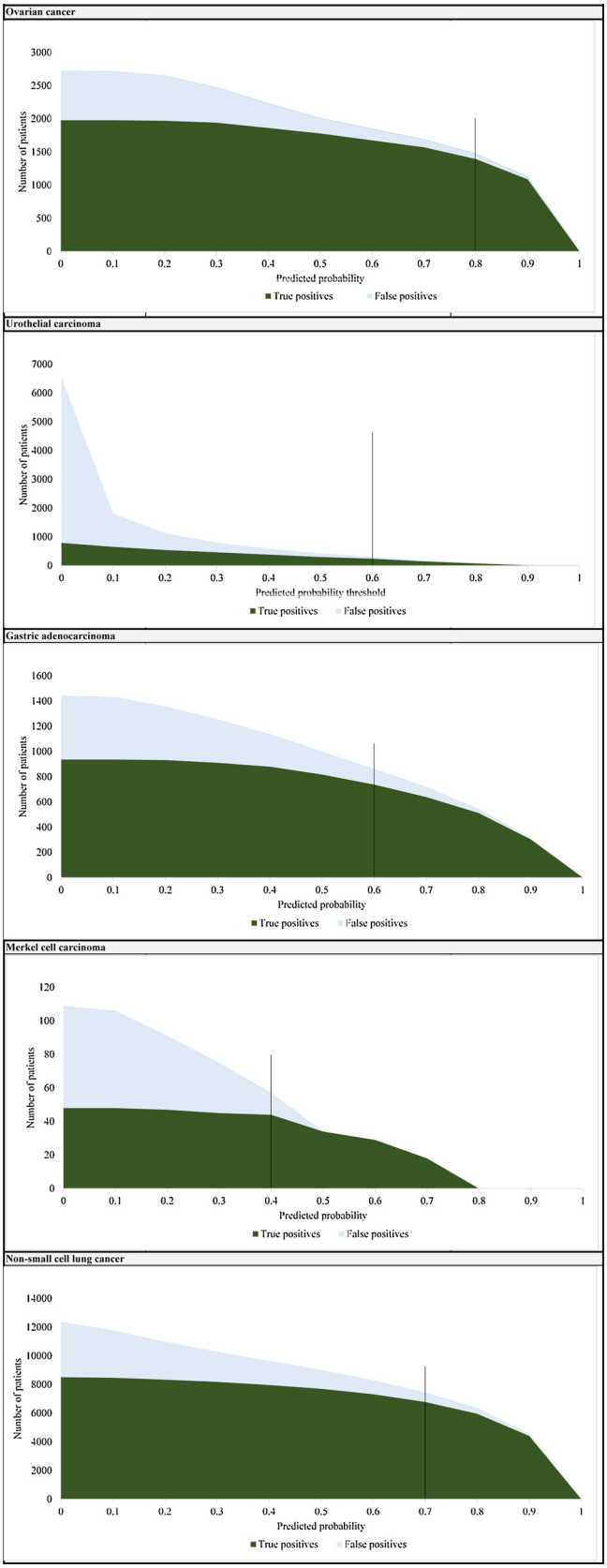

Fig. 2. True and false positive advanced cancer cases by site and algorithm predictive value threshold.

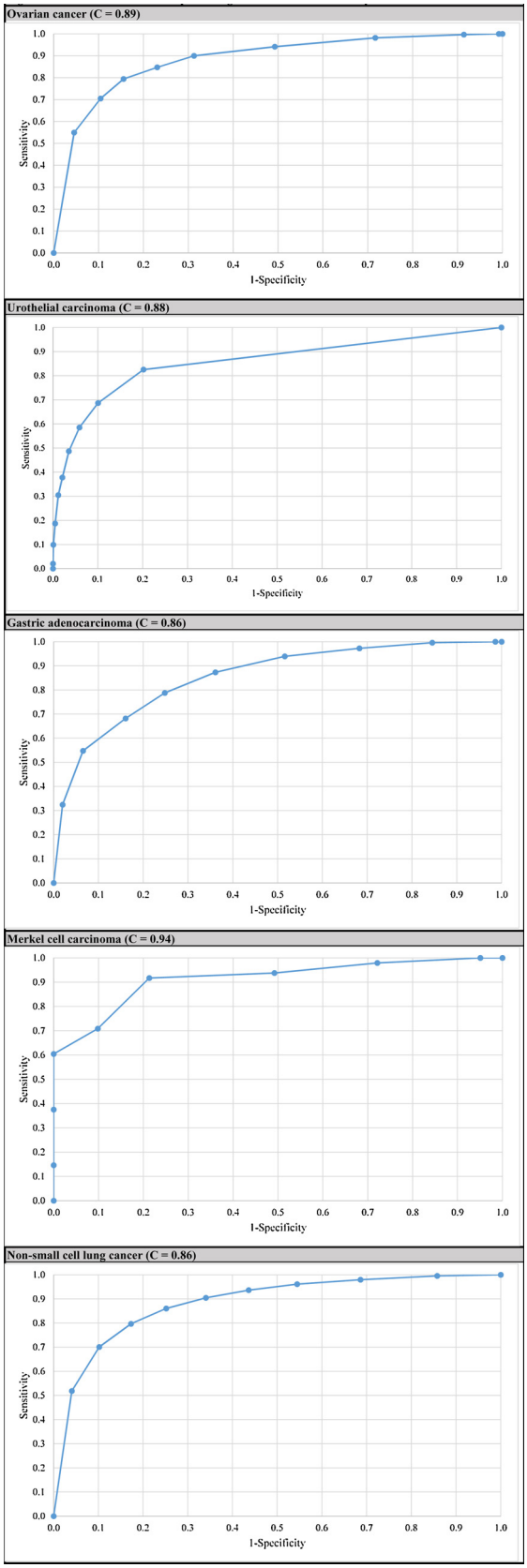

Fig. 3. ROC curves for models predicting advanced cancer status by site.

The LASSO regression model performed moderately at lower thresholds and well at probability thresholds above 0.30, with a cross-validated PPV ranging from 0.69 to 0.97 (Table 2). Of the 102 variables assigned a non-zero weight by LASSO regression, the strongest predictors of confirmed metastatic/advanced NSCLC included treatment with vinblastine (β = 2.60), treatment with crizotinib (β = 1.45), and pneumonectomy (β = 1.28). The strongest predictors of confirmed non-metastatic/advanced NSCLC included at least one surgical procedure (β = -1.22), treatment with antineoplastic antibiotics (β = -1.13), and treatment with chemotherapy rescue/antidote agents (β = -1.10 [Supplemental Table 2]).

The number of false positives decreased from 3308 at a predicted probability threshold of 0.1–157 at a predicted probability of 0.9 (Fig. 2). A threshold of 0.70 was selected, with a PPV of 0.91 (95% CI 0.90 – 0.92) and a sensitivity of 80% (Table 2). This model had a c-statistic of 0.89 indicating a high degree of discrimination between true and false cases of metastatic/advanced NSCLC (Fig. 3).

Applying the predictive models to the HIRD population identified 7473 patients with metastatic/advanced ovarian cancer, 2155 patients with metastatic/advanced urothelial cancer, 6941 patients with metastatic/advanced gastric adenocarcinoma, 795 patients with metastatic/advanced MCC, and 42,490 patients with metastatic/advanced NSCLC.

4. Discussion

Diagnostic codes do not indicate cancer staging, limiting the utility of healthcare claims databases in oncology research. To address this limitation, a subset of cancer patients from the HIRD claims data were linked to cancer registries and HIRE-O data that provided clinical details unavailable in claims, including primary site and stage. This sample of individuals with confirmed cancer stage was used to develop and validate predictive models. By applying the predictive models to the large claims database, we accurately identified large populations of patients classified as having metastatic/advanced cancer stage even though this information is missing in the source claims data.

Case definitions that are based on predictive models augment a priori knowledge with empirically-derived predictors, and offer investigators flexibility in selecting a threshold to define case status. The high discrimination in these algorithms (i.e., c-statistics > 0.85) is due in part to the large number of covariates included in the predictive models, as well as range of coefficients that capture the strength and direction of the association of each predictor with confirmed cancer stage.

Most algorithms used in administrative claims data specify Boolean joins between variables with a binary outcome defining case status. An advantage of using predictive models is that the researcher can easily modify the probability threshold at which a subject is classified as a case in order to balance false negative and false positive errors to minimize misclassification bias to suit the needs of a particular study [12]. For example, a study seeking to include only true cases would favor a high PPV algorithm, noting the cost of missed cases. Likewise, to conduct comparative analyses with ratio measures (e.g., rate ratio), it might be desirable to set a threshold that corresponds to a high PPV (including few non-cases) following the principle that a non-differential loss of sensitivity with the highest possible specificity yields an unbiased estimate on the ratio measure [13]. To estimate incidence, however, a highly sensitive algorithm (i.e., lower threshold) could be used and the estimated incidence corrected for the false positive error rate. Because different study goals motivate different trade-offs between sensitivity and PPV, flexibility in balancing errors is valuable.

Some limitations of the predictive models are important to consider. As with any case identifying algorithm, it was developed in a particular time and setting, and algorithms can perform differently when applied in a different time period or place. For instance, performance of the models developed here might degrade as clinical practice patterns change and coding systems evolve. While this concern is not unique to model-based algorithms, it may be particularly important for the algorithms derived in this study. Various treatments were frequently represented as covariates, and these models may need to be updated as treatments change.

As with other algorithms, these models have potential to be applied in other data sources that contain the same administrative data elements (e.g., longitudinal capture of diagnoses and procedures from healthcare settings and prescription medications), however, data captured in each environment may not be comparable. Even where the underlying data elements included in a database are the same, completeness and accuracy of recording certain diagnoses or performing certain procedures could vary by setting, limiting algorithm transportability. Also, the HIRD, as with many claims databases, is missing considerable data on race and lifestyle factors such as smoking, body mass index, and alcohol consumption. Were these parameters available and used in model development in a different environment, the resulting models could differ and accuracy could be improved.

The HIRD is a dynamic population, and the length of time for which data are available varies from person to person. Because administrative claims capture only those diagnoses that are recorded at the healthcare encounter and those with longer follow-up typically have more recorded encounters, it is possible that these individuals are more accurately classified by our algorithms. In so far as longer duration of follow-up after the first diagnosis of cancer was a weak indicator of early stage for the ovarian, gastric and non-small cell lung cancer sites and was not selected by LASSO models for Merkel cell or urothelial carcinoma, it appears that duration of follow-up did not have a substantial influence in our data.

To the extent that actual cases of metastatic/advanced cancer never receive a diagnosis code for their cancer, the initial inclusion criteria would not capture all metastatic/advanced cancer cases in the source claims population, and the sensitivity of the model would be overestimated if applied to the entire HIRD. We believe this to be unlikely, but it is difficult to demonstrate. Further, patients linked to cancer registries or HIRE-O were not random samples. We do not have reason to believe that cancer patients in the HIRD included in this study would have differed from those in other US claims data sources on predictors for advanced cancer status, however testing the accuracy of these models using other data sources would help ascertain their transportability.

Finally, while generally considered reliable, cancer registries and HIRE-O data are not necessarily fully accurate in assessing metastatic/advanced cancer status. Misclassification could occur due to error in entry or inaccurate clinical assessment of stage at the time of recording. Despite high concordance between medical records and HIRE-O and registry data, a small degree of misclassification of metastatic/advanced cancer status is still possible. While further assessment of portability across data sources, times and places is necessary to maximize the utility of these predictive models, as it is for any case identifying algorithm, predictive models allow for substantial refinement of metastatic/advanced cancer case definitions in data sources where cancer stage is absent.

Funding

This work was sponsored by Merck KGaA, Darmstadt, Germany and Pfizer Inc., New York, NY, USA and is part of an alliance between Merck KGaA, Darmstadt, Germany and Pfizer Inc., New York, NY, USA. HealthCore, Inc. received funding from the sponsors to conduct the study.

Disclaimers

Cancer incidence data used in these analyses were obtained from the Ohio Cancer Incidence Surveillance System (OCISS), Ohio Department of Health (ODH), a cancer registry partially supported by

the National Program of Cancer Registries at the Centers for Disease Control and Prevention (CDC) through Cooperative Agreement Number NU58DP006284. Use of these data does not imply that ODH or CDC agrees or disagrees with the analyses, interpretations or conclusions in this report.

Author statement

Per the guidelines adopted by the International Committee of Medical Journal Editors ("Uniform requirements for manuscripts submitted to biomedical journals", authorship should be based on:

1 substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data;
2 drafting the article or revising it critically for important intellectual content; and
3 final approval of the version to be published.

The below authors have each contributed to the manuscript in all three capacities as outlined above.

Authors: Daina B Esposito, Leo Russo, Dina Oksen, RuihuaYin, Vibha CA Desai, Jennifer G Lyons, Patrice Verpillat, Jose L Peñalvo, Francois-Xavier Lamy, Stephan Lanes.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.canep.2019.05.006.

References

[1] B.L. Nordstrom, et al., Validation of claims algorithms for progression to metastatic cancer in patients with breast, non-small cell lung, and colorectal cancer, Front. Oncol. 6 (2016) 18.
[2] B.L. Nordstrom, et al., Identification of metastatic cancer in claims data, Pharmacoepidemiol. Drug Saf. 21 (Suppl. 2) (2012) 21–28.
[3] G.S. Cooper, et al., The utility of Medicare claims data for measuring cancer stage, Med. Care 37 (7) (1999) 706–711.
[4] J.L. Whyte, et al., An evaluation of algorithms for identifying metastatic breast, lung, or colorectal cancer in administrative claims data, Med. Care 53 (7) (2015) e49–57.
[5] N. Chawla, et al., Limited validity of diagnosis codes in Medicare claims for identifying cancer metastases and inferring stage, Ann. Epidemiol. 24 (9) (2014) 666–672 672.e1-2.
[6] A.L. Mahar, et al., Validating an algorithm to identify metastatic gastric cancer in the absence of routinely collected TNM staging data, BMC Health Serv. Res. 18 (1) (2018) 309.
[7] M.J. Hassett, et al., Detecting lung and colorectal cancer recurrence using structured clinical/administrative data to enable outcomes research and population health management, Med. Care 55 (12) (2017) e88–e98.
[8] J. Barron, J. Malin, A. Ganetsky, D.M. Kern, B. Wu, O. Tunceli, V.J. Wiley, R.A. Quimbo, M. Fisch, J. Singer, R. Mamtani, Validity of clinical information for a novel cancer care quality program, J. Clin. Oncol. 35 (Suppl. (15)) (2016) e18003.
[9] Institute, N.C, DCCPS, S.R. Program, Seer Research Data Record Description Cases Diagnosed In 1973-2009*, (2011) Surveillance, Epidemiology, and End Results (SEER) Program, (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER 9 Regs Research Data.
[10] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. 58 (1) (1996) 267–288 Series B.
[11] J.H. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Softw. 33 (1) (2010) 1–22.
[12] J. Chubak, G. Pocobelli, N.S. Weiss, Tradeoffs between accuracy measures for electronic health care data algorithms, J. Clin. Epidemiol. 65 (3) (2012) 343–349 e2.
[13] C. Poole, Exception to the rule about nondifferential misclassification ([Abstract]), Am. J. Epidemiol. 122 (1985).