original report

Predicting Renal Cancer Recurrence: Defining Limitations of Existing Prognostic Models With Prospective Trial-Based Validation

Andres F. Correa, MD[1]; Opeyemi Jegede, MPH[2]; Naomi B. Haas, MD[3]; Keith T. Flaherty, MD[4]; Michael R. Pins, MD[5]; Edward M. Messing, MD[6]; Judith Manola, MS[2]; Christopher G. Wood, MD[7]; Christopher J. Kane, MD[8]; Michael A.S. Jewett, MD[9]; Janice P. Dutcher, MD[10]; Robert S. DiPaola, MD[11]; Michael A. Carducci, MD[12]; and Robert G. Uzzo, MD[1]

abstract

PURPOSE To validate currently used recurrence prediction models for renal cell carcinoma (RCC) by using prospective data from the ASSURE (ECOG-ACRIN E2805; Adjuvant Sorafenib or Sunitinib for Unfavorable Renal Carcinoma) adjuvant trial.

PATIENTS AND METHODS Eight RCC recurrence models (University of California at Los Angeles Integrated Staging System [UISS]; Stage, Size, Grade, and Necrosis [SSIGN]; Leibovich; Kattan; Memorial Sloan Kettering Cancer Center [MSKCC]; Yaycioglu; Karakiewicz; and Cindolo) were selected on the basis of their use in clinical practice and clinical trial designs. These models along with the TNM staging system were validated using 1,647 patients with resected localized high-grade or locally advanced disease (≥ pT1b grade 3 and 4/pTanyN1Mo) from the ASSURE cohort. The predictive performance of the model was quantified by assessing its discriminatory and calibration abilities.

RESULTS Prospective validation of predictive and prognostic models for localized RCC showed a substantial decrease in each of the predictive abilities of the model compared with their original and externally validated discriminatory estimates. Among the models, the SSIGN score performed best (0.688; 95% CI, 0.686 to 0.689), and the UISS model performed worst (0.556; 95% CI, 0.555 to 0.557). Compared with the 2002 TNM staging system (C-index, 0.60), most models only marginally outperformed standard staging. Importantly, all models, including TNM, demonstrated statistically significant variability in their predictive ability over time and were most useful within the first 2 years after diagnosis.

CONCLUSION In RCC, as in many other solid malignancies, clinicians rely on retrospective prediction tools to guide patient care and clinical trial selection and largely overestimate their predictive abilities. We used prospective collected adjuvant trial data to validate existing RCC prediction models and demonstrate a sharp decrease in the predictive ability of all models compared with their previous retrospective validations. Accordingly, we recommend prospective validation of any predictive model before implementing it into clinical practice and clinical trial design.

J Clin Oncol 37:2062-2071. © 2019 by American Society of Clinical Oncology

ASSOCIATED CONTENT

Data Supplement

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

Accepted on April 23, 2019 and published at jco.org on June 19, 2019: DOI https://doi.org/10.1200/JCO.19.00107

INTRODUCTION

In the management of cancer, accurate predictive tools are essential for effective patient counseling, surveillance, development of adjuvant strategies, and clinical trial design. In the absence of reliable biomarkers, clinicians rely primarily on the combination of stage, grade, and histology to predict oncologic events. For nearly eight decades, the TNM system has occupied the central role in risk prediction and therefore communication and resource allocation.[1] Unfortunately, in the era of personalized medicine, its ability to accurately predict individual patient oncologic outcomes is limited.[2] Along this continuum, advanced statistical methods have led to the development of predictive models based largely on retrospective, categorical data that seek to improve individualized recurrence predictions.

In the management of renal cell carcinoma (RCC), adoption of these prognostic models has become central to patient counseling, clinical guideline development, and adjuvant trial design. Currently, eight prognostic algorithms and nomograms that were developed from retrospective single-institutional experience over the last three decades are widely used for predicting the risk of relapse in RCC.[3-10] Each model considers clinical and/or pathologic variables but differs with regard to the number and type of covariates, tool properties (nomogram or prognostic categories), and end points (overall survival, cancer-specific survival, and recurrence-free survival). Most importantly, these models are all retrospective, and although more than 30 external validations have been published that included more than 37,000 patients,

the performance of these models has never been tested by using prospective data.

The development of these models preceded the clinical availability of the first reliably successful systemic therapies for metastatic RCC, including antiangiogenic targeted and immunologic therapies. Because there have been efforts to use effective systemic therapies earlier in the disease process, these models became central to the design of adjuvant RCC trials. Specifically, the eligibility for each of the 11 largest RCC adjuvant trials (Data Supplement) that have been completed, or are currently accruing, rely on the predictive/prognostic abilities of these models, which are based on the retrospective recurrence data of approximately 815 patients (Table 1) with localized disease. Considered another way, 13,000 patients are or will be enrolled into these trials (collectively costing hundreds of millions of dollars) on the basis of the predictive ability of these models. An often overlooked fact is that thousands of other patients were or will be excluded from these trials based on the strengths or weakness of the three principle predictive RCC models (University of California at Los Angeles Integrated Staging System [UISS][5]; Leibovich[10]; and Stage, Size, Grade and Necrosis [SSIGN] score[6]).

To better understand the implications of using retrospective predictive models for adjuvant clinical trial design, we validated the performance of the eight most commonly used RCC predictive models by using prospective data from the largest, placebo-controlled, adjuvant trial conducted to date, the ASSURE (ECOG-ACRIN E2805; Adjuvant Sorafenib or Sunitinib for Unfavorable Renal Carcinoma) trial.[11] To our knowledge, this is the first time prospective, highly annotated, and centrally reviewed data have been used to validate any of the current RCC predictive models.

PATIENTS AND METHODS

Study Population

The study population consists of patients recruited for the ASSURE trial,[11] the first and largest adjuvant trial assessing the benefit of targeted therapy (sunitinib or sorafenib compared with placebo) in patients with intermediate- or high-risk localized kidney cancer. Eligibility criteria have been previously published[11] and are summarized in the Data Supplement. Central pathology review allowed the variables central to all models (grade, stage, and histology) to be standardized. Demographic variables were collected as a standard component of trial eligibility (symptoms at presentation and performance status). Patients were assessed every three cycles (18 weeks) by computed tomography or magnetic resonance imaging scans for recurrence during the first year. Patients were then observed using scans and laboratory and clinical assessments every 6 months for another year, then once per year until disease recurrence or through 10 years. Accrual to the ASSURE trial was completed in September 2010 with recurrence data collected through December 18, 2017.

Description of Prognostic Models

A total of 10 models[3-10,12,13] developed to predict RCC recurrence were identified from the medical literature. Eight models[3-10] (Table 1) were selected for analysis on the basis of their use in clinical trial design, popularity in the clinical setting, and previous validation across independent cohorts. For reference purposes, the discriminatory ability of the three most recent editions (5th, 6th, and 7th) of the kidney cancer TNM staging system were also validated. ASSURE used the 6th TNM staging edition for inclusion criteria; another TNM staging classification was compiled using the 5th and 7th TNM editions as a result of the excellent annotation of pathologic variables in ASSURE.

Outcome Variables

The primary outcome measure in the ASSURE trial[11] was disease-free survival (DFS), defined as the time from random assignment to recurrence, development of second primary cancer, or death as a result of any cause. Patients alive without disease recurrence at the time of analysis were censored on the date of last contact. Although the clinical outcome for some of the models differed and included outcomes other than those specified in ASSURE (overall survival, cancer-specific survival, metastasis-free survival), the duration of disease follow-up allowed the evaluation of these secondary end points.

Validation of Existing Prognostic Models Using ASSURE Data

Model validation was performed according to transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines.[14] Each model was validated by using its respective prediction outcome. The intention was to validate each model for discrimination and calibration (whenever possible, given the parameters provided in the original publication).

Discrimination was measured using the concordance statistic (C-index).[15] The C-index for survival outcomes is an extension of the area under the receiver operating characteristic curve for binary outcomes; its value ranges from 0.5 through 1. A C-index value of 1 indicates perfect discrimination, and 0.5 indicates a model no better than a fair coin toss. The C-index was assessed by using the approaches of both Harrell et al[16] and Uno et al.[17] The method of Harrell et al provides an overall measure of discrimination, and the method of Uno et al estimates discrimination from baseline to a specific time point. The linear predictor used in the assessment of discrimination was calculated as the sum of the product of regression coefficients (reported for each model) and variables.[18,19] For models that did not report threshold values risk categories (Kattan and MSKCC), we decided to use the sum of points obtained from the nomogram presented in the respective

TABLE 1. Validated C-Indices for Common Prognostic Models

Model	UISS[5]	SSIGN[6]	Leibovich[10]	MSKCC*[4]	Kattan*[3]	Yaycioglu[9]	Karakiewicz[7]	Cindolo[8]
Type	K-M survival analysis	Algorithm	Algorithm	Nomogram	Nomogram	Formula	Nomogram	Formula
Outcome	OS	CSS	MFS	RFS	RFS	RFS	CSS	RFS
Time period	1989-1999	1970-1999	1970-2000	1989-2002	1989-1998	1990-1999	1984-2006	1987-2003
Original No. of patients/No. of patients with localized disease (TanyN0M0)	814/468†	1,801/1,417†	1,671/1,605†	701	601	296	2,474/1,967†	660
No. of events for patients with localized disease (TanyN0M0)	116‡	NR	479	72	66	38	NR	110
Histology	ccRCC/Pap/Chromo	ccRCC	ccRCC	ccRCC	ccRCC/Pap/Chromo	ccRCC/Pap/Chromo	ccRCC/Pap/Chromo	ccRCC/Pap/Chromo
Inclusion criteria	RNx/PNx, pTanyN0-2M0-1	RNx/PNx, pTanyN0-2M0-1	RNx, pTanyN0-2M0	RNx/PNx, pT1-3bN0M0	RNx/PNx, pT1-3N0M0	RNx/PNx, pT1-3cN0M0	RNx/PNx, pT1-3N0-2M0-1	RNx/PNx, pT1-3N0M0
Model variables	3 Factors: 1997 TNM, Fuhrman grade, ECOG PS	6 Factors: 1997 TNM, pN+, pM+, tumor size, Fuhrman grade, tumor necrosis	5 Factors: 1997 TNM, pN+, tumor size, Fuhrman grade, tumor necrosis	5 Factors: 2002 TNM, tumor size, Fuhrman grade, necrosis, symptoms at presentation	4 Factors: 1997 TNM, tumor size, histology, symptoms at presentation	2 Factors: clinical tumor size, symptoms at presentation	6 Factors: 2002 TNM, age, sex, cm+, tumor size, symptoms at presentation	2 Factors: clinical tumor size, symptoms at presentation
Median follow-up (years)	2.5	9.7	5.4	2.7	3.3	4	12	3.5
Reported C-index	0.73	0.84	0.82	0.82	0.74	0.65	0.86	0.67
Validation								
No. of validations	4	5	3	2	7	3	2	3
No. of patients	7,594	6,909	2,530	1,043	6,306	3,274	6,605	4,045
Validation C-index	0.64-0.86	0.76-0.88	0.70-0.82	0.79-0.82	0.27-0.84	0.63-0.70	0.78-0.88	0.63-0.75
ASSURE C-index (95% CI)	0.556 (0.555 to 0.557)	0.69 (0.686 to 0.689)	0.625 (0.623 to 0.626)	0.652 (0.650 to 0.653)	0.622 (0.621 to 0.623)	0.587 (0.585 to 0.588)	0.617 (0.616 to 0.619)	0.632 (0.630 to 0.633)

Abbreviations: ccRCC, clear cell renal cell carcinoma; Chromo, chromophobe; CSS, cancer-specific survival; ECOG PS, Eastern Cooperative Oncology Group performance status; K-M, Kaplan-Meier; MFS, metastasis-free survival; MSKCC, Memorial Sloan Kettering Cancer Center; NR, not reported; OS, overall survival; Pap, papillary; PNx, partial nephrectomy; RFS, recurrence-free survival; RNx, radical nephrectomy; SSIGN, Stage, Size, Grade and Necrosis score; UISS, University of California at Los Angeles Integrated Staging System.

*New model was fitted by using ASSURE data and thus does not constitute classic model validation.

†Number of patients with localized disease in the model (TanyN0M0).

‡Calculated from the percent disease-free survival estimates at 5-year follow-up.

manuscript as a risk score (linear predictor) for each patient, and we calculated C-index appropriately using the sum of the points.

Model calibration was assessed by using calibration plots. The plot depicts predicted versus observed 5-year recurrence-free survival probabilities; a good calibration is indicated by a close alignment of predicted and observed probability estimates along the diagonal.

RESULTS

In all, 1,647 patients in the ASSURE cohort met inclusion criteria for analysis (Data Supplement). A detailed clinicopathologic description of the cohort can be found in the Data Supplement. The mean tumor size was 8.6 cm (± 3.4 cm), with more than half the patients (62.9%) having broadly defined tumor-related symptoms at presentation. The majority of the tumors analyzed were clear cell RCC (80.6%), 64.7% were categorized as high grade (Fuhrman grade 3 or 4), 42.2% of the patients presented with evidence of tumor necrosis, and 8.0% presented with nodal involvement. At a median follow-up of 7.85 years, 785 (47.7%) disease-specific, 440 (20.5%) overall, 338 (20.5%) RCC-specific, 436 (26.5%) metastatic, and 741 (45.0%) recurrence events were observed.

The distribution of risk categories for each of the evaluated predictive models is shown in Table 2. Consistent with the eligibility criteria for the ASSURE trial, most models stratified patients into intermediate- or high-risk categories of an adverse oncologic outcome. The estimated 1-, 3-, 5-, 7-, and 9-year survival estimates for each model stratified by risk category were calculated on the basis of data from the ASSURE trial (Table 2).

A summary of existing predictive models is presented in Table 1. The UISS,[5] Karakiewicz,[7] Yaycioglu,[9] and Cindolo[8] models assessed preoperative characteristics, whereas the Kattan,[3] MSKCC,[4] Leibovich[10] and SSIGN[6] models focused on postresection variables. The originally published and reported C-index for the models ranged from a low of 0.58 in the Yaycioglu[9] model to a high of 0.86 in the Karakiewicz[7] model. The Kattan[3] (seven external validations) and SSIGN[6] (five external validations) models have had the most published external validations, with C-indices ranging from 0.76 to 0.88. Application of each model to the 1,647 patients from the ASSURE trial[11] showed a significant decrease in the discriminatory ability of each model, with calculated C-indices ranging from 0.556 (95% CI, 0.555 to 0.557) in the UISS model[5] to 0.688 (95% CI, 0.686 to 0.689) in the SSIGN score[6] (Table 1). Every model performed well below their originally reported C-index, and nearly all performed below the calculated C-index in the previous external validations. By using the standard TNM (2002) staging as a reference (C-index, 0.60), six of the eight validated models were found to (marginally) outperform the predictive accuracy of the standard TNM staging criteria, with the UISS[5] and Yaycioglu[9] models being notable exceptions (Fig 1A). Furthermore, the predictive ability (C-index) of each model was found to be highly variable over time (Fig 1B), reaching peak discriminatory ability at or before 2 years of follow-up.

A calibration evaluation was performed for the MSKCC[4] and Kattan[3] models because they were published in nomogram form. Although the Karakiewicz[7] model was published in nomogram form, we chose to forgo a calibration evaluation because the model is under-specified in this analysis because it excluded metastatic patients, and an approach to calibrate it on the basis of risk groupings would have been equivocal. For the remainder of the models, a baseline hazard function was lacking, which is required to perform an accurate calibration validation of Cox prognostic models.[18] Overall, the MSKCC[4] model had better 5-year progression-free survival prediction probabilities than the Kattan[3] model (Fig 2). The Kattan[3] model tended to significantly overestimate the risk of recurrence compared with the observed events in the ASSURE trial. The MSKCC[4] model significantly underestimated the recurrence rates of high-risk patients, although it accurately predicted 5-year progression-free survival for low- and intermediate-risk individuals.

DISCUSSION

The ability to predict future oncologic events has broad significance to patients and clinicians. In the absence of reliable and validated biomarkers, clinical and pathologic parameters remain the primary variables in communicating prognosis, implementing surveillance strategies, recommending adjuvant therapies, and designing clinical trials. As institutional databases and big data efforts have emerged, the last 20 years have seen a surge in the development, validation, and implementation of presumably more robust prognostic models aimed at providing increasingly accurate and individualized assessments. In RCC, adoption of these prognostic models have rapidly become the standard for patient risk stratification in adjuvant trial design. Conflicting results from the RCC adjuvant trials[11,19-21] in which these predictive models were used have led to questions about their accuracy and generalizability, which has set a precedent for the careful adoption of similar models in other malignancies. Here, we assessed the performance of existing kidney cancer prediction models by using highly annotated, prospective data from the ASSURE trial to reconcile the results observed in current RCC adjuvant trials and demonstrate the inherent limitations of prognostic models upon which significant resources are leveraged.

Application of prospective data with central pathology review to the eight most commonly used kidney cancer prediction models demonstrates that they all significantly underperform their original and externally validated discriminatory estimates. When testing the C-index using the

TABLE 2. Patient Risk Distribution by Model With 1-, 3-, 5-, 7-, and 9-Year Survival Estimate

	Cohort (n = 1,647)		Time (years)*									
			1		3		5		7		9	
Model and Scores†	No.	%	Survival Estimate (%)	95% CI	Survival Estimate (%)	95% CI	Survival Estimate (%)	95% CI	Survival Estimate (%)	95% CI	Survival Estimate (%)	95% CI
UISS (OS)[5]												
Low risk, I	6	1	—		—		—		—		—	
Intermediate risk, II	1379	85	97.9	97.0 to 98.6	90.1	88.3 to 91.6	82.8	80.6 to 84.7	77.3	74.9 to 79.6	71.1	68.1 to 73.9
High risk			92.7	88.5 to 95.4	78.0	72.1 to 82.9	67.9	61.3 to 73.6	61.1	54.2 to 67.3	48.3	40.0 to 56.0
III	182	11										
IV	49	3										
V	9	1										
SSIGN (CSS)[6]			—		—		—		—		—	
0-2	78	6										
3-4	437	33	99.8	98.6 to 100	97.1	95.2 to 98.3	93.6	91.0 to 95.5	91.2	88.1 to 93.4	85.8	81.2 to 89.4
5-6	425	33	99.5	98.0 to 99.9	91.9	88.8 to 94.2	84.6	80.6 to 87.9	77.3	72.5 to 81.3	73.0	67.3 to 77.9
7-9	351	27	93.9	90.8 to 95.9	80.1	75.5 to 83.9	71.0	65.9 to 75.5	66.7	61.3 to 71.5	59.7	53.4 to 65.5
≥ 10	12	1										
Leibovich (MFS)[10]												
Low (0-2)	8	1	—		—		—		—		—	
Intermediate (3-5)	777	59	94.6	92.7 to 96.0	84.4	81.6 to 86.9	79.6	76.5 to 82.4	75.2	71.8 to 78.3	71.0	67.1 to 74.6
High (≥ 6)	522	40	89.7	86.7 to 92.1	69.3	65.0 to 73.3	61.8	57.2 to 66.1	55.9	51.0 to 60.5	54.6	49.5 to 59.4
Yaycioglu (RFS)[9]												
Low	865	61	90.5	88.3 to 92.3	76.6	73.5 to 79.3	69.5	66.2 to 72.6	64.4	61.0 to 67.7	58.1	54.1 to 61.8
High	563	39	83.4	80.0 to 86.3	62.9	58.7 to 66.9	55.4	51.1 to 59.6	48.4	43.9 to 52.7	45.0	40.2 to 49.6
Cindolo (RFS)[8]												
Low	385	28	92.3	89.1 to 94.6	79.7	75.3 to 83.5	73.8	68.9 to 78.0	70.0	64.9 to 74.4	61.4	55.3 to 66.9
High	1009	72	86.3	84.0 to 88.3	68.1	65.0 to 70.9	60.4	57.2 to 63.4	53.6	50.3 to 56.8	50.0	46.5 to 53.5

(continued on following page)

TABLE 2. Patient Risk Distribution by Model With 1-, 3-, 5-, 7-, and 9-Year Survival Estimate (continued)

	Cohort (n = 1,647)		Time (years)*									
			1		3		5		7		9	
Model and Scores†	No.	%	Survival Estimate (%)	95% CI	Survival Estimate (%)	95% CI	Survival Estimate (%)	95% CI	Survival Estimate (%)	95% CI	Survival Estimate (%)	95% CI
MSKCC points (RFS)‡[4]												
Low to < 120 (Q1)	285	24	95.7	92.5 to 97.5	86.4	81.8 to 90.0	79.8	74.5 to 84.2	75.9	70.1 to 80.6	71.1	64.6 to 76.6
148 to < 120 (Q2)	303	25	92.1	88.4 to 94.7	76.7	71.4 to 81.2	68.2	62.4 to 73.3	61.8	55.7 to 67.3	55.2	48.5 to 61.4
148 to 176 (Q3)	308	25	86.1	81.6 to 89.6	66.0	60.2 to 71.1	58.1	52.2 to 63.6	47.5	41.3 to 53.4	40.9	34.1 to 47.6
176 to high	313	26	77.0	71.9 to 81.3	48.8	43.1 to 54.2	40.1	34.6 to 45.6	33.5	28.2 to 38.9	31.0	25.4 to 36.8
Kattan points (RFS)‡§[3]												
Low to < 80 (Q1)	315	21	92.7	89.2 to 95.2	86.7	82.3 to 90.0	79.0	73.9 to 83.2	76.0	70.6 to 80.5	70.4	64.2 to 75.8
120 to < 100 (Q2)	263	18	92.2	88.1 to 94.9	78.8	73.1 to 83.3	70.6	64.4 to 75.9	64.6	58.1 to 70.4	59.1	51.9 to 65.6
148 to < 120 (Q3)	522	36	89.1	86.0 to 91.5	68.0	63.7 to 71.9	61.5	57.0 to 65.7	54.7	50.0 to 59.1	49.3	44.2 to 54.3
120 to high	371	25	80.3	75.8 to 84.0	58.0	52.8 to 62.9	49.5	44.2 to 54.5	41.4	36.1 to 46.6	38.2	32.4 to 43.9
Karakiewicz (CSS)‖[7]												
Low	709	44	98.7	97.5 to 99.3	94.9	93.0 to 96.3	89.8	87.2 to 91.9	88.3	85.5 to 90.6	84.5	80.9 to 87.4
Intermediate	745	46	96.8	95.2 to 97.9	87.2	84.5 to 89.5	81.6	78.4 to 84.3	75.5	72.0 to 78.6	71.0	66.8 to 74.7
High	156	10	94.8	89.9 to 97.4	76.5	69.0 to 82.5	69.0	60.9 to 75.7	63.2	54.7 to 70.5	55.9	45.3 to 65.2

NOTE. Dashes indicate calculation not performed because of the low number of patients.

Abbreviations: CSS, cancer-specific survival; MFS, metastasis-free survival; MSKCC, Memorial Sloan Kettering Cancer Center; OS, overall survival; Q1, first quartile; RFS, recurrence-free survival; SSIGN, Stage, Size, Grade and Necrosis score; UISS, University of California at Los Angeles Integrated Staging System.

*The specific survival outcome for each model is included in parentheses.

†The inclusion criteria listed in Table 1 have been applied to models with overall No. of patients below 1,647.

‡Arbitrary (quartile) categories of points obtained from provided nomogram.

§Proportion of patients in each category do not seem to be equal because of the level of precision of points.

‖Three-category risk group was created from a multivariable model linear predictor.

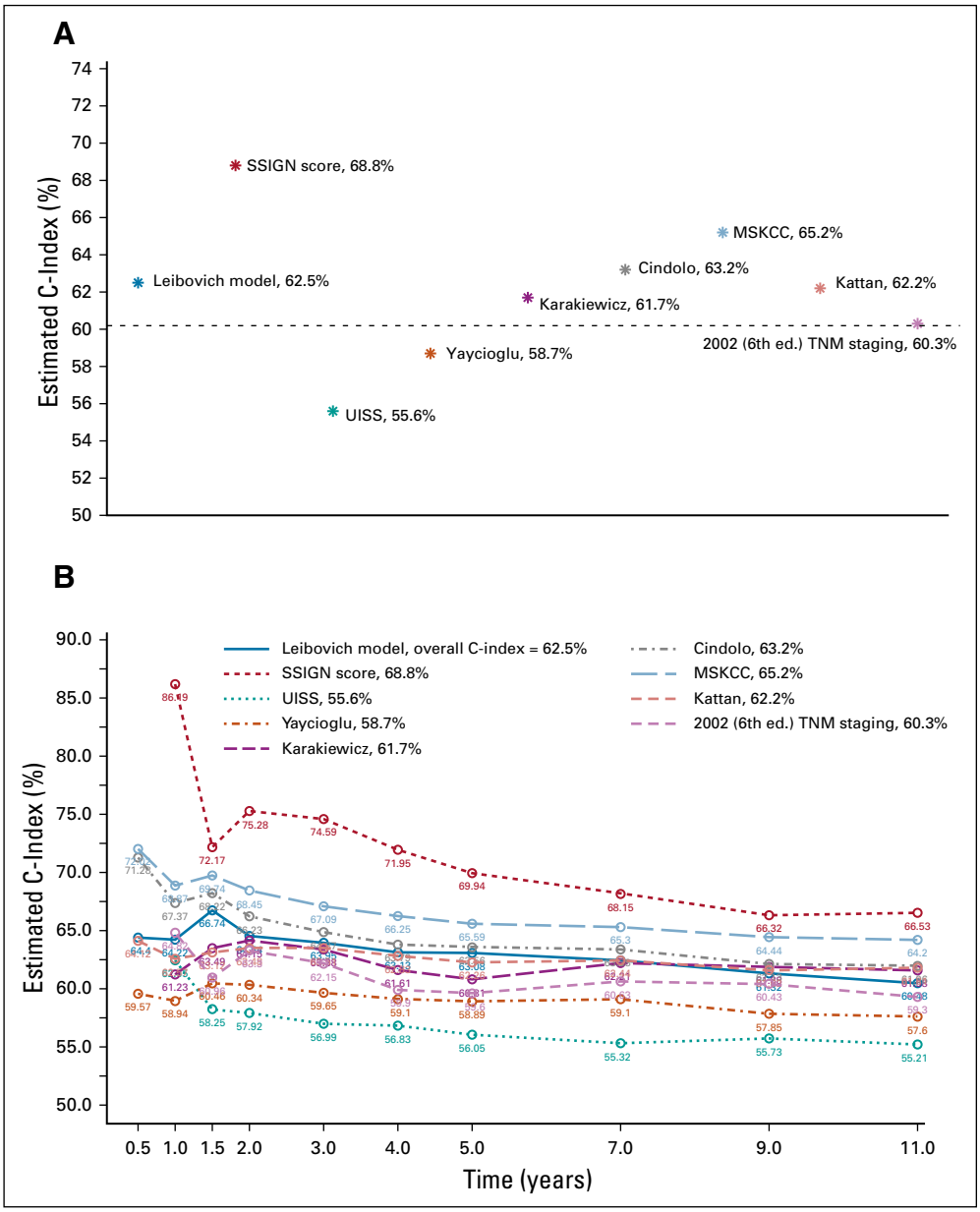

FIG 1. The C-index estimate for each of the eight models validated by using data from the ASSURE trial. (A) Overall C-index estimate; dashed line delineates the TNM C-index threshold. (B) The C-index estimate over time. MSKCC, Memorial Sloan Kettering Cancer Center; SSIGN, State, Size, Grade, Necrosis score; UISS, University of California at Los Angeles Integrated Staging System.

primary end point of each tool, among adjuvant relevant models, the SSIGN[6] score performed the best (0.688; 95% CI, 0.686 to 0.689), and the UISS[4] model performed the worst (0.556; 95% CI, 0.555 to 0.557). Models that included tumor biology factors (MSKCC,[4] Leibovich,[10] and SSIGN[6]) tended to outperform those that included patient symptoms at presentation (UISS,[5] Cindolo,[8] Karakiewicz,[7] Yaycioglu[9]). In comparison with the 2002 TNM staging[22] (Fig 1), the validated models marginally outperform standard staging, with two of the models (UISS[5] and Yaycioglu[9]) demonstrating decreased predictive accuracy. Interestingly, the updates made to the TNM staging over the last three TNM versions have only modestly improved the predictive accuracy of the system (Data Supplement) when validated using prospective clinical trial data. Here we have provided the observed survival estimates for each model at 1, 3, 5, 7, and 9 years, which can be referenced for those currently using or planning to use the models validated herein.

All models (including the three TNM staging models) showed significant variability in their prediction accuracies over time, reaching their peak predictive ability within the first 2 years after diagnosis (Fig 1B; Data Supplement). This time-dependent degradation is a likely representation of the biologic forces contributing to the dichotomy of cancer recurrences (early *v* late), with early recurrences being highly influenced by tumor biology (ie, easily predicted by tumor-centric models), whereas late recurrences are the result of a complex interplay between host factors (immune surveillance) and tumor biology. Furthermore, the skewed distribution of postresection RCC events (70% of recurrences occur within the first 2 years)[23,24] is likely a contributing factor to the predictive degradation phenomenon

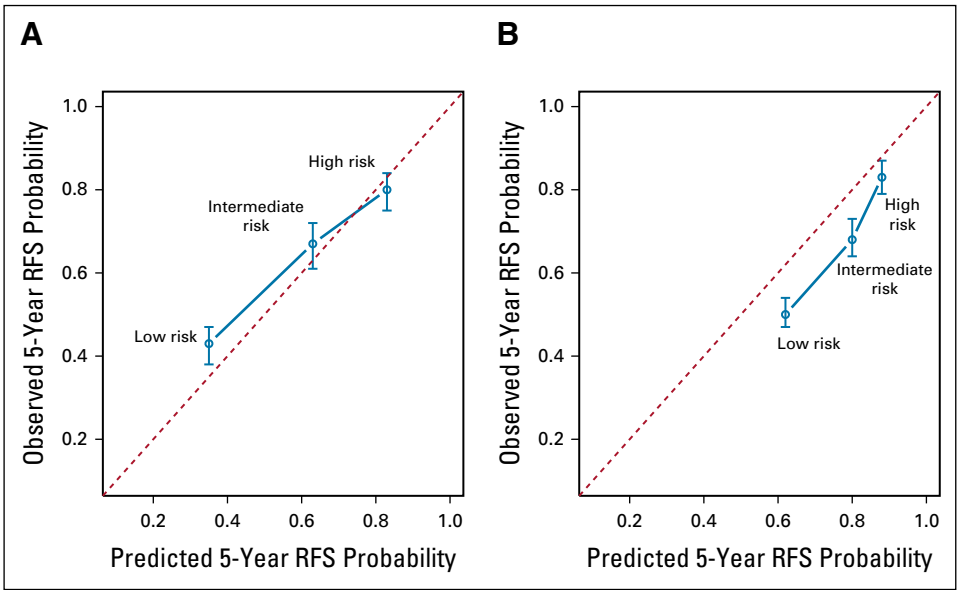

FIG 2. Calibration plot of the Memorial Sloan Kettering Cancer Center (MSKCC) and Kattan nomograms for the entire cohort of 1,647 patients. Three risk categories were estimated on the basis of the model outputs (low, intermediate, and high). Estimates above the dashed line represent underestimates; those below the dashed line represent overestimates. RFS, recurrence-free survival.

because recurrences become less frequent and unevenly distributed with the passage of time.

To date, all of the reported, pending, and currently recruiting adjuvant trials (Data Supplement) rely on these models for statistical design and patient eligibility. The inherent risk of this cannot be understated, whereas when these models are validated using prospective data, most of them seem marginally more discriminative than a coin flip. This level of discriminatory ability coincides with the observed clinical enrichment in these trials in which only 40% to 50% of untreated patients were found to develop a clinical recurrence[11,19,20] compared with the expected clinical enrichment of 60% to 70%.[11,19,20] Interestingly, the use of the superior SSIGN score (C-index, 0.688) by the PROTECT (A Study to Evaluate Pazopanib as an Adjuvant Treatment for Localized Renal Cell Carcinoma) trial (2-year DFS, 65.7%) did not translate into an improved clinical enrichment compared with the ASSURE trial (2-year DFS, 63.7%) and the S-TRAC (A Clinical Trial Comparing Efficacy and Safety of Sunitinib Versus Placebo for the Treatment of Patients at High Risk of Recurrent Renal Cell Cancer) trial (2-year DFS, 66.9%), which used the lesser performing UISS (C-index, 0.556) model. The marginal improvement in clinical enrichment observed by using these models is likely related to their marginal superiority to the standard TNM staging system. The current evaluation shows that inclusion of biologic tumor factors (tumor necrosis) in the model tends to improve their predictive ability compared with those that include less objectively measured patient factors (symptoms at presentation). Inclusion of further biologic risk (genetic and immune signatures) factors in these models should be given high consideration. The recurrence score,[25] which incorporates a set of RCC driver mutations, has been shown to be an independent predictor of recurrence after adjustment for the Leibovich score. The design of the PROSPER (A Phase 3 Randomized Study Comparing Perioperative Nivolumab vs. Observation in Patients With Renal Cell Carcinoma Undergoing Nephrectomy) trial (neoadjuvant *v* adjuvant checkpoint inhibitors) has a unique potential for identifying immune signatures which can be then added to the existing and future models to further improve their predictive accuracy.

Our analysis raises significant concerns regarding overreliance on retrospective prognostic models for cancer prognostication and resource allocation. To date, prognostic nomograms have been created for almost all solid malignancies, and most of them are based on retrospective data.[26] There are multiple sources of error when retrospective data are used for model development, including differences in data collection techniques, lack of standardization, differences in reporting, shifting practices, changes in collection methods over time, evolution in surgical techniques, and lack of a centralized pathologic review. Likewise, the use of retrospective data to externally validate prognostic models poses a similar risk of collection and confirmation bias. Furthermore, the publication of external validation data often serves to enhance the credibility of the model and further entrenches its use in the literature and clinical practice.

The unregulated use of clinical prognostic models in clinical practice has led the American Joint Committee on Cancer (AJCC) to set standard criteria for model selection.[27] None of the models validated here meet AJCC criteria required for model use (Data Supplement), mainly because of their lack of validation in contemporary data sets. Furthermore, McGinn and colleagues[28] have provided a comprehensive guide for the evaluation of clinical prediction tools. In their guide, they introduced a hierarchy of evidence for the adoption of a prediction model, in which

three of the four levels of evidence require model validation in a prospective sample (Levels 1 to 3), with models validated in large prospective multicenter cohorts providing a higher level of evidence (Levels 1 and 2). The retrospective validations were deemed to provide the lowest level of evidence (Level 4). The vast amount of high-quality data that has been and is being generated in multiple adjuvant RCC trials should provide a fruitful ground for the development and validation of high-impact prognostic models that may one day achieve the clinical enrichment that was hoped for with the current models. Importantly, tumor-centric models should be developed that focus on early recurrence because their predictive ability tends to degrade over time. We believe that identification of late recurrences will require more complex algorithms that can model host and tumor interactions or that use novel genomic technologies such as liquid biopsies.[29]

This study provides the highest level of validation to date for the most commonly used RCC prediction models, some currently used in the design of costly adjuvant clinical trials. The validation was carried out in data collected for a large prospective multicenter trial, with centralized evaluation of clinical and pathologic variables along with standardized reporting of outcome measures (Level 2). The validation presented is not devoid of limitations. First, the cohort and survival data used for this validation originated from an adjuvant trial in which two thirds of the patients received one of two targeted agents (sunitinib or sorafenib) for a 12-month period. Although the overall effect of adjuvant therapy on the trial outcome measure (DFS) was not significant,[11] minor treatment effects could have affected the observed outcomes, thus confounding the validation of these treatment-naïve prognostic models. A sensitivity analysis using the placebo group was performed that showed only minor differences in the calculated C-index (Data Supplement), confirming the low potential for treatment effect bias in this validation. A calibration test was not performed in most of the evaluated models because of the lack of a published baseline hazard function for all of the non-nomogram models. Although a calibration test is important for model evaluation in regard to prediction accuracy, discrimination (the ability to select those at risk) is the most important measure of a model used for clinical trial eligibility and thus was the focus of this analysis.

In conclusion, we have demonstrated that the most commonly used risk evaluation tools for localized RCC have significant and underappreciated limitations. The initial reports of their predictive ability and subsequent external validations are hampered by multiple unmeasurable variations in retrospective data collection and therefore their performance, when measured against robustly annotated prospective data, is significantly diminished. Therefore, we recommend the use of prospective data to provide final validation of prognostic models before their adoption.

AFFILIATIONS

[1]Fox Chase Cancer Center, Philadelphia, PA
[2]Dana-Farber Cancer Institute, Boston, MA
[3]University of Pennsylvania, Philadelphia, PA
[4]Massachusetts General Hospital, Boston, MA
[5]Advocate Lutheran General Hospital, Park Ridge, IL
[6]University of Rochester, Rochester, NY
[7]MD Anderson Cancer Center, Houston, TX
[8]University of California-San Diego, La Jolla, CA
[9]University of Toronto, Toronto, ON, Canada
[10]Cancer Research Foundation, New York, NY
[11]University of Kentucky College of Medicine, Lexington, KY
[12]Johns Hopkins Hospital, Baltimore, MD

CORRESPONDING AUTHOR

Robert G. Uzzo, MD, Department of Surgical Oncology, Fox Chase Cancer Center, Temple University Health System, 333 Cottman Ave, Philadelphia, PA 19111; e-mail: robert.uzzo@fccc.edu.

SUPPORT

Supported by Grants No. CA180820, CA180794, CA180867, CA180858, CA180888, CA180821, CA180863 from the National Cancer Institute, National Institutes of Health, and Grant No. 704970 from the Canadian Cancer Society.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST AND DATA AVAILABILITY STATEMENT

Disclosures provided by the authors and data availability statement (if applicable) are available with this article at DOI https://doi.org/10.1200/JCO.19.00107.

AUTHOR CONTRIBUTIONS

Conception and design: Andres F. Correa, Opeyemi Jegede, Naomi B. Haas, Keith T. Flaherty, Michael R. Pins, Judith Manola, Christopher G. Wood, Michael A. Carducci, Robert G. Uzzo

Administrative support: Janice P. Dutcher, Robert G. Uzzo

Provision of study materials or patients: Naomi B. Haas, Michael A.S. Jewett, Michael A. Carducci, Robert G. Uzzo

Collection and assembly of data: Opeyemi Jegede, Naomi B. Haas, Michael R. Pins, Edward M. Messing, Judith Manola, Christopher J. Kane, Michael A.S. Jewett, Robert S. DiPaoloa

Data analysis and interpretation: Andres F. Correa, Opeyemi Jegede, Naomi B. Haas, Keith T. Flaherty, Christopher G. Wood, Janice P. Dutcher, Robert S. DiPaoloa, Michael A. Carducci, Robert G. Uzzo

Manuscript writing: All authors

Final approval of manuscript: All authors

Accountable for all aspects of the work: All authors

ACKNOWLEDGMENT

We thank the manuscript reviewers for their insights and constructive comments. This study was coordinated by the ECOG-ACRIN Cancer Research Group (Peter J. O'Dwyer, MD and Mitchell D. Schnall, MD, PhD, Group Co-Chairs).

REFERENCES

1. Denoix PF: [Nomenclature and classification of cancers based on an atlas] [Article in Undetermined Language]. Acta Unio Int Contra Cancrum 9:769-771, 1953
2. Amin MB, Greene FL, Edge SB, et al: The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA Cancer J Clin 67:93-99, 2017
3. Kattan MW, Reuter V, Motzer RJ, et al: A postoperative prognostic nomogram for renal cell carcinoma. J Urol 166:63-67, 2001
4. Sorbellini M, Kattan MW, Snyder ME, et al: A postoperative prognostic nomogram predicting recurrence for patients with conventional clear cell renal cell carcinoma. J Urol 173:48-51, 2005
5. Zisman A, Pantuck AJ, Wieder J, et al: Risk group assessment and clinical outcome algorithm to predict the natural history of patients with surgically resected renal cell carcinoma. J Clin Oncol 20:4559-4566, 2002
6. Frank I, Blute ML, Cheville JC, et al: An outcome prediction model for patients with clear cell renal cell carcinoma treated with radical nephrectomy based on tumor stage, size, grade and necrosis: The SSIGN score. J Urol 168:2395-2400, 2002
7. Karakiewicz PI, Briganti A, Chun FK, et al: Multi-institutional validation of a new renal cancer-specific survival nomogram. J Clin Oncol 25:1316-1322, 2007
8. Cindolo L, de la Taille A, Messina G, et al: A preoperative clinical prognostic model for non-metastatic renal cell carcinoma. BJU Int 92:901-905, 2003
9. Yaycioglu O, Roberts WW, Chan T, et al: Prognostic assessment of nonmetastatic renal cell carcinoma: A clinically based model. Urology 58:141-145, 2001
10. Leibovich BC, Blute ML, Cheville JC, et al: Prediction of progression after radical nephrectomy for patients with clear cell renal cell carcinoma: A stratification tool for prospective clinical trials. Cancer 97:1663-1671, 2003
11. Haas NB, Manola J, Uzzo RG, et al: Adjuvant sunitinib or sorafenib for high-risk, non-metastatic renal-cell carcinoma (ECOG-ACRIN E2805): A double-blind, placebo-controlled, randomised, phase 3 trial. Lancet 387:2008-2016, 2016
12. Thompson RH, Leibovich BC, Lohse CM, et al: Dynamic outcome prediction in patients with clear cell renal cell carcinoma treated with radical nephrectomy: The D-SSIGN score. J Urol 177:477-480, 2007
13. Karakiewicz PI, Suardi N, Capitanio U, et al: Conditional survival predictions after nephrectomy for renal cell carcinoma. J Urol 182:2607-2612, 2009
14. Moons KG, Altman DG, Reitsma JB, et al: New guideline for the reporting of studies developing, validating, or updating a multivariable clinical prediction model: The TRIPOD statement. Adv Anat Pathol 22:303-305, 2015
15. Caetano SJ, Sonpavde G, Pond GR: C-statistic: A brief explanation of its construction, interpretation and limitations. Eur J Cancer 90:130-132, 2018
16. Harrell FE Jr, Lee KL, Mark DB: Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 15:361-387, 1996
17. Uno H, Cai T, Pencina MJ, et al: On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med 30: 1105-1117, 2011
18. Royston P, Altman DG: External validation of a Cox prognostic model: Principles and methods. BMC Med Res Methodol 13:33, 2013
19. Ravaud A, Motzer RJ, Pandha HS, et al: Adjuvant Sunitinib in High-Risk Renal-Cell Carcinoma after Nephrectomy. N Engl J Med 375:2246-2254, 2016
20. Motzer RJ, Haas NB, Donskov F, et al: Randomized phase III trial of adjuvant pazopanib versus placebo after nephrectomy in patients with localized or locally advanced renal cell carcinoma. J Clin Oncol 35:3916-3923, 2017
21. Pfizer/News: Pfizer provides update on phase 3 trial of axitinib as adjuvant treatment for patients at high risk of renal cell carcinoma recurrence after surgery. April 10, 2018. https://www.pfizer.com/news/press-release/press-release-detail/pfizer_provides_update_on_phase_3_trial_of_axitinib_as_adjuvant_treatment_for_patients_at_high_risk_of_renal_cell_carcinoma_recurrence_after_surgery
22. Greene FL, Page DL, Fleming ID, et al: Kidney, in Edge SB, Byrd DR, Compton CC, et al (eds): AJCC Cancer Staging Manual. New York, NY, Springer, 2002, pp. 323-328
23. McNichols DW, Segura JW, DeWeerd JH: Renal cell carcinoma: Long-term survival and late recurrence. J Urol 126:17-23, 1981
24. Ljungberg B, Alamdari FI, Rasmuson T, et al: Follow-up guidelines for nonmetastatic renal cell carcinoma based on the occurrence of metastases after radical nephrectomy. BJU Int 84:405-411, 1999
25. Rini B, Goddard A, Knezevic D, et al: A 16-gene assay to predict recurrence after surgery in localised renal cell carcinoma: Development and validation studies. Lancet Oncol 16:676-685, 2015
26. Mallett S, Royston P, Waters R, et al: Reporting performance of prognostic models in cancer: a review. BMC Med 8:21, 2010
27. Kattan MW, Hess KR, Amin MB, et al: American Joint Committee on Cancer acceptance criteria for inclusion of risk models for individualized prognosis in the practice of precision medicine. CA Cancer J Clin 66:370-374, 2016
28. Guyatt G, Rennie D, Meade MO, et al (eds): Clinical Prediction Rules, in Guyatt G, Rennie D, Meade MO, et al (eds): Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice (ed 3). New York, NY, McGraw-Hill Education, 2015
29. Sparano J, O'Neill A, Alpaugh K, et al: Association of circulating tumor cells with late recurrence of estrogen receptor-positive breast cancer: A secondary analysis of a randomized clinical trial. JAMA Oncol 4:1700-1706, 2018

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Predicting Renal Cancer Recurrence: Defining Limitations of Existing Prognostic Models With Prospective Trial-Based Validation

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. 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/jco/site/ifc.

Naomi B. Haas
Consulting or Advisory Role: Pfizer, Merck Sharp & Dohme
Expert Testimony: Eli Lilly (I)

Keith T. Flaherty
Stock and Other Ownership Interests: Clovis Oncology, Loxo Oncology, X4 Pharmaceuticals, Strata Oncology, PIC Therapeutics, Fount Therapeutics, Shattuck Labs, Apricity Health, Oncoceutics, FogPharma, Tvardi Therapeutics, Vivid Biosciences, Checkmate Pharmaceuticals
Consulting or Advisory Role: Novartis, Genentech, Merck, Eli Lilly, Amgen, Sanofi, Oncoceutics, Bristol-Myers Squibb, Adaptimmune Therapeutics, Aeglea Biotherapeutics, Loxo Oncology, Roche, Asana BioSciences, Incyte, Shattuck Labs, Tolero Pharmaceuticals, Array BioPharma, FogPharma, Neon Therapeutics, Tvardi Therapeutics, Takeda Pharmaceuticals, Verastem, Boston Biomedical, Pierre Fabre, Cell Medica, Debiopharm Group
Research Funding: Novartis, Sanofi
Travel, Accommodations, Expenses: Pierre Fabre, Debiopharm Group

Michael R. Pins
Speakers' Bureau: Genentech, Merck

Christopher G. Wood
Leadership: Kidney Cancer Association
Honoraria: Pfizer, Merck, Argos Therapeutics
Speakers' Bureau: Pfizer, Argos Therapeutics
Research Funding: Argos Therapeutics, Pfizer
Travel, Accommodations, Expenses: Argos Therapeutics, Merck

Christopher J. Kane
Stock and Other Ownership Interests: SNP Bio

Michael A.S. Jewett
Stock and Other Ownership Interests: Theralase Technologies
Honoraria: Pfizer, Theralase Technologies
Patents, Royalties, Other Intellectual Property: Patent application filed for a new RFA device

Janice P. Dutcher
Consulting or Advisory Role: Prometheus Laboratories, Bristol-Myers Squibb/Medarex, TRACON Pharmaceuticals, Nektar Therapeutics, Amgen, Merck, Eisai, Iovance Biotherapeutics
Speakers' Bureau: Prometheus Laboratories
Travel, Accommodations, Expenses: Prometheus Laboratories

Michael A. Carducci
Consulting or Advisory Role: Astellas Pharma, AbbVie, Genentech, Pfizer, Foundation Medicine
Research Funding: Bristol-Myers Squibb (Inst), Pfizer (Inst), AstraZeneca (Inst), Gilead Sciences (Inst), EMD Serono (Inst), eFFECTOR Therapeutics (Inst)

Robert G. Uzzo
Consulting or Advisory Role: Johnson & Johnson, Pfizer, Genentech
Speakers' Bureau: Janssen Oncology
Research Funding: Novartis

No other potential conflicts of interest were reported.