External Validation of a Nomogram Using RENAL Nephrometry Score to Predict High Grade Renal Cell Carcinoma

Hong-Kai Wang,* Yao Zhu,* Xu-Dong Yao, Shi-Lin Zhang, Bo Dai, Hai-Liang Zhang, Yi-Jun Shen, Chao-Fu Wang and Ding-Wei Ye†

From the Departments of Urology (HKW, YZ, XDY, SLZ, BD, HLZ, YJS, DWY) and Pathology (CFW), Fudan University Shanghai Cancer Center and Department of Oncology, Shanghai Medical College, Fudan University (HKW, YZ, XDY, SLZ, BD, HLZ, YJS, CFW, DWY), Shanghai, People’s Republic of China

Purpose: A novel nomogram using the RENAL ([R]adius maximal diameter in cm, [E]xophytic/endophytic properties, [N]earness of the tumor to the collecting system or sinus in mm, [A]nterior/Posterior, [L]ocation relative to the polar lines and [H]ilar) nephrometry score was developed to predict high grade renal cell carcinoma. It showed good performance in internal evaluation. We externally validated the prediction model.

Materials and Methods: We identified a cohort of 391 Chinese patients in whom renal cell carcinoma was surgically resected at our institution from 2008 to 2011. Fuhrman grade was reviewed by an experienced genitourinary pathologist and radiological images were independently assessed by 2 senior urologists. Using a 2-tiered system high grade disease was defined as Fuhrman grade III/IV. The statistical performance of the prediction model was evaluated by discrimination, calibration and clinical usefulness.

Results: Of the 391 patients 45.5% were considered to have high grade tumors. External validation of the nomogram revealed an AUC of 0.73. The calibration plot showed that the predicted probability of high grade disease had concordance comparable to the observed frequency. On decision curve analysis the prediction model provided a superior net benefit and reduction at a greater than 20% probability threshold.

Conclusions: We confirm the predictive value of the nomogram using the RENAL nephrometry score to identify high grade renal cell carcinoma in an independent cohort. Further research is required to evaluate its performance using a head-to-head comparison with renal biopsy results.

Key Words: kidney; carcinoma, renal cell; nomograms; validation studies; decision support techniques

Abbreviations and Acronyms

FG = Fuhrman grade
RCC = renal cell carcinoma

Submitted for publication September 21, 2011.
Study received institutional review board approval.
Supported by National Natural Science Foundation of China (Project 81001131).
* Equal study contribution.
† Correspondence: Department of Urology, Fudan University Shanghai Cancer Center, No. 270 Dong’an Rd., Shanghai, 200032, People’s Republic of China (telephone: 86-21-64175590-2807; FAX: 86-21-64434556; e-mail: dwye@shca.org.cn).
prediction models by including FG as an essential part.\textsuperscript{4,5}

In the era of target therapy high FG is strongly associated with poor survival in patients with metastatic RCC treated with sunitinib.\textsuperscript{6} Thus, estimating tumor grade before treatment substantially aids in prognosis estimation, patient consultation and clinical decision making.

Currently percutaneous renal biopsy is the first choice for pretreatment estimation of the pathological characteristics of a renal mass. Although the technique has 97\% overall accuracy for a malignant diagnosis, the concordance rate of FG is less promising.\textsuperscript{7} Even in contemporary series the reported accuracy of tumor grade is 43\% to 97\% (mean of flipping a coin (0.5).\textsuperscript{11}

Currently percutaneous renal biopsy is the first choice for pretreatment estimation of the pathological characteristics of a renal mass. Although the technique has 97\% overall accuracy for a malignant diagnosis, the concordance rate of FG is less promising.\textsuperscript{7} Even in contemporary series the reported accuracy of tumor grade is 43\% to 97\% (mean of flipping a coin (0.5).\textsuperscript{11} Several studies suggest that biopsy sensitivity and negative predictive value are associated with tumor size.\textsuperscript{8,9} Rybicki et al found highest sensitivity and negative predictive value (97\% and 89\%) for masses 4 to 6 cm while sensitivity and negative predictive value sharply decreased for masses less than 3 cm (84\% and 60\%) and greater than 6 cm (87\% and 44\%, respectively).\textsuperscript{9} To assist in precisely estimating FG prediction models were developed using clinical variables such as patient age, gender, symptom and tumor size.\textsuperscript{10,11} However, the predictive accuracy of these tools was close to that of flipping a coin (0.5).\textsuperscript{11}

Recently Kutikov et al constructed a novel nomogram that incorporated the RENAL nephrometry score to predict the probability of high grade disease.\textsuperscript{12} When adding anatomical features, the model showed an AUC of 0.73 in the development cohort. Before these encouraging findings can be used in routine clinical practice they must be validated in an independent external cohort. Thus, we assessed the validity of the nomogram of Kutikov et al to predict high grade RCC.

MATERIALS AND METHODS

Patient Population and Risk Calculation

The records of patients who underwent renal surgery at Fudan University Shanghai Cancer Center from January 2008 to January 2011 were collected from the electronic medical database. After excluding those with missing information we identified 391 consecutive cases of a single enhancing renal mass and a pathological diagnosis of RCC. Clinical characteristics, pathological slides and cross-sectional images were retrieved for all enrolled patients.

FG was reviewed by an experienced genitourinary pathologist (CFW) and any discrepancy with a former report was resolved by a conference review. The 4-tiered FG system was simplified into a 2-tiered classification, as in the report by Kutikov et al.\textsuperscript{12} FG I and II were considered low grade, and FG III and IV were considered high grade. Two senior urologists (YZ and HKW) blinded to pathological outcome independently evaluated the radiological images and assigned the 6 components (R, E, N, A, L and H) of the nephrometry score.\textsuperscript{13} The Appendix shows details of the scoring system.

Interobserver variation was measured by Cohen $\kappa$ statistical analysis. The values of $\kappa$ were 0.82, 0.76, 0.65, 0.90, 0.69, and 0.67 for the R, E, N, A, L and H parts of the score, respectively. For all scoring components $p < 0.05$, indicating that the estimated $\kappa$ was not a result of chance. For inconsistent scores the final decisions were made by jointly reviewing the images. The predicted risk of high grade RCC was determined from an online risk calculator.\textsuperscript{12} This retrospective study was approved by the institutional review board and written informed consent was provided by each patient.

Statistical Analysis

We report descriptive statistics of the validation and development cohorts. Statistical performance of the prediction model was assessed by discrimination, calibration and clinical usefulness.\textsuperscript{14} The AUC was used to determine discrimination ability between patients with high vs low grade disease. The AUC 95\% CI was calculated by bootstrapping using 200 resamples. Calibration refers to the agreement between observed and predicted outcomes. The extent of overestimation/underestimation relative to observed and predicted rates was explored graphically using calibration plots. Clinical usefulness was assessed by decision curve analysis\textsuperscript{15} to estimate a net benefit for prediction models by summing the benefits (positive findings) and subtracting the harms (false-positive findings). The latter were weighted by a factor related to the relative harm of missed high grade cancer vs an overrated tumor. The interpretation of a decision curve was that the model with the highest net benefit at a particular threshold probability should be chosen. For all analyses 2-sided $p < 0.05$ was considered significant. Statistical analysis was done using R 2.13.0.

RESULTS

The table lists the clinicopathological characteristics of the 391 patients enrolled in the validation data set and the 453 in the original developmental data set.\textsuperscript{12} Median age at surgery was 55 (range 15 to 84) in the validation set and 60 years (range 25 to 89) in the developmental set. Median tumor size was 4.2 (range 1.2 to 13) and 4 cm (range 0.7 to 25), respectively. There were significant differences between the initial cohort and the validation cohort in histological subtype (see table). In the validation cohort a lower rate of nonclear cell carcinoma was identified than in the initial cohort (12.0\% vs 28.0\%). The distribution of other characteristics was comparable between the 2 data sets.

External validation of the nomogram revealed an AUC of 0.73 (95\% CI 0.68–0.78, fig. 1). The calibration plot showed that the predicted probability of high grade disease had concordance comparable to that of the observed frequency with most predictions within a 5\% margin of error (fig. 2). The Hosmer-
Lemeshow goodness of fit test revealed $p = 0.11$ (chi-square statistic 14.39), indicating nonsignificant miscalibration. Decision curve analysis revealed that the prediction model provided superior net benefit and reduction with a probability threshold of greater than 20% (fig. 3).

**DISCUSSION**

The RENAL nephrometry scoring system was developed by Kutikov et al to quantitatively assess the anatomical features of renal masses. The scoring system is associated with surgical respectability and with pathological characteristics. Using components of the RENAL nephrometry scoring system Kutikov et al developed a nomogram to predict malignant and high grade disease. In their report the AUC was 0.76 for the cancer model and 0.73 for the grade model. Compared to the cancer model the grade model is more attractive since FG determination is still inaccurate in renal biopsy specimens.

Application of the novel nomogram was limited by the lack of external validation. In our study we used a Chinese cohort of 391 patients to validate the grade model. We confirmed that the prediction model can accurately estimate the probability of high grade disease.

The nomogram AUC of 0.73 in the validation cohort confirmed good discrimination ability between high and low grade disease. The achieved predictive accuracy of the nomogram in the validation cohort was the same as the statistic in the original data set. These promising results suggested a robust prediction model that sufficiently captured informative predictive factors.

This may also have been caused by a smaller proportion of nonclear cell carcinoma in our patient population than in the development cohort (12.0% vs 28.0%). It is well known that the probability of high grade disease varies by RCC histological subtype. Descriptive data on an international data set of 4,063 RCC cases revealed that 44.2% of clear cell carcinomas were high grade while 33.1% of papillary and 32% of chromophobe carcinomas were classified as high grade. In contrast, sarcomatoid and collecting duct cell carcinomas are highly aggressive. de Peralta-Venturina et al found that 86 of 101 sarcomatoid cell carcinomas were high grade. Similarly 32 of 40 collecting duct carcinomas were high grade in the series by Karakiewicz et al. Thus, there may be variations in the prognostic effect of the RENAL nephrometry score between clear cell and other renal tumor subtypes. Due to the few nonclear cell carcinoma cases (47) in our study further evaluation of model performance in this subgroup was not attempted. The hypothesis must be further tested in large multicenter studies with a large enough sample size.

The calibration plot showed good agreement between predicted probability and observed frequency. The nomogram underestimated the probability of high grade disease by greater than 5% only in cases within a predicted probability range of 40% to 45%. With a wide range of favorable concordance the nomogram could provide great value for appropriately interpreting renal biopsy findings.

In addition to discrimination and calibration, we further evaluated the clinical usefulness of the nomogram to facilitate decision making on further

<table>
<thead>
<tr>
<th>Clinicopathological characteristics of validation and developmental data sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. Validation (%)</td>
</tr>
<tr>
<td><strong>Gender:</strong></td>
</tr>
<tr>
<td>M</td>
</tr>
<tr>
<td>F</td>
</tr>
<tr>
<td><strong>Tumor histology:</strong></td>
</tr>
<tr>
<td>Clear cell RCC</td>
</tr>
<tr>
<td>Papillary RCC</td>
</tr>
<tr>
<td>Chromophobe RCC</td>
</tr>
<tr>
<td>Sarcomatoid RCC</td>
</tr>
<tr>
<td>Other malignant type</td>
</tr>
<tr>
<td><strong>Tumor stage:</strong></td>
</tr>
<tr>
<td>High</td>
</tr>
<tr>
<td>Low</td>
</tr>
<tr>
<td><strong>T stage:</strong></td>
</tr>
<tr>
<td>T1</td>
</tr>
<tr>
<td>T2</td>
</tr>
<tr>
<td>T3</td>
</tr>
<tr>
<td>T4</td>
</tr>
<tr>
<td><strong>N stage:</strong></td>
</tr>
<tr>
<td>Nx/N0</td>
</tr>
<tr>
<td>N+</td>
</tr>
<tr>
<td><strong>M stage:</strong></td>
</tr>
<tr>
<td>Mx/M0</td>
</tr>
<tr>
<td>M1</td>
</tr>
</tbody>
</table>

* In report by Kutikov et al gender, age and tumor size data included 72 benign renal tumors with FG not specified in 10 patients.
interventions. The prediction tool showed better net benefit and reduction at a threshold probability of 20% or greater. Today the management of a small renal mass involves a balance of oncological risk, comorbidity, life expectancy and renal function.\textsuperscript{7} Thus, the nomogram may be especially helpful for treatment selection in poor surgical candidates who may need to be classified at higher risk for aggressive cancer to avoid active surveillance.

Using components of the RENAL nephrometry score the nomogram of Kutikov et al\textsuperscript{12} showed satisfactory statistical performance in our cohort. However, there is still debate about the prognostic value of some elements in the model, especially the E and H scores.\textsuperscript{21} Although previous reports suggested that exophytic renal tumors are more likely to be of low grade, most assessed tumors were small.\textsuperscript{22,23} The correlation between E score and FG may be decreased in large renal tumors since a proportion of these masses tends to grow outward and may still be classified with an E of 1. Also, Lipke et al found that exophytic renal tumors are more likely to be of a papillary subtype.\textsuperscript{22} In the series by Kutikov et al papillary RCC also showed a higher probability of being of the exophytic type than of the clear cell or chromophobe subtype.\textsuperscript{12} Thus, the prognostic significance of E score may probably be influenced by histological subtype composition.

Although few groups have addressed the impact of renal hilar involvement on high grade disease, several analyzed the prognostic effect of renal sinus fat invasion.\textsuperscript{24} Recently Bertini et al found that localized RCC with sinus fat invasion showed a higher probability of high grade disease than cases with perinephric fat invasion (74.2\% vs 48.6\%).\textsuperscript{25}

Fully independent validation studies performed by other investigators at other sites are more con-
vincing to evaluate the generalizability of prediction models. However, validation samples should derive from different but plausibly related populations. Except for the distribution of histological subtypes the baseline characteristics of our patients were comparable to those of the original population. The 88% proportion of clear cell RCC in our study is higher than the 72% rate in the initial report but similar to that in other large series. Clear cell carcinoma accounted for 84.5% of cases in a multicenter study of 2,030 Chinese patients with RCC. Another international study revealed that 87.7% of 4,063 cases were classified as clear cell RCC.

To our knowledge the underlying mechanisms contributing to the different prevalence of histological subtypes are poorly understood. Smoking history was found in 35% of our patients, similar to contemporary Western series. A cytogenetic study of RCC from Southeast Asian patients also revealed no specific alterations different from those of their Western counterparts. Nevertheless, integrating new predictors associated with RCC histological subtypes might further improve nomogram accuracy and generalizability.

Our study had several limitations. 1) Enrolled patients were retrospectively collected at a single tertiary center and so results may have been influenced by selection bias. 2) The end point of the original and the validation study was FG, which may show interobserver variation. However, the drawback was minimized by a second review and by a 2-tiered system. Using a simplified FG system is supported by higher agreement among pathologists as well as by evidence of predictive accuracy similar to that of the classic system. In a large multicenter study of 5,453 patients with RCC similar performance was observed after adding 2, 3 and 4-tiered FG systems in a multivariate model to predict cancer specific survival. 3) The study population could have been larger. However, the number of participants was sufficient for an external validation study and fulfilled event per variable criteria. Despite these limitations we thoroughly assessed the statistical performance of the nomogram of Kutikov et al and confirmed the validity of the prediction model in an independent Asian cohort.

**CONCLUSIONS**

We confirmed the predictive value of the nomogram using the RENAL nephrometry score to identify high grade RCC in an independent cohort. Further research is required to evaluate its performance using a head-to-head comparison with renal biopsy results.

**APPENDIX**

**Description of RENAL Nephrometry Score**

<table>
<thead>
<tr>
<th>Description</th>
<th>1 Point</th>
<th>2 Points</th>
<th>3 Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>(R)adius (maximal diameter in cm)</td>
<td>4 or Less</td>
<td>Greater than 4-less than 7</td>
<td>7 or Greater</td>
</tr>
<tr>
<td>(E)xophytic/endophytic properties</td>
<td>50% or Greater</td>
<td>Less than 50%</td>
<td>Entirely endophytic</td>
</tr>
<tr>
<td>(N)earness of tumor to collecting system or sinus (mm)</td>
<td>7 or Greater</td>
<td>Greater than 4-less than 7</td>
<td>4 or Less</td>
</tr>
<tr>
<td>(A)nterior/Posterior</td>
<td>Entirely above upper or below lower polar line</td>
<td>Lesion crosses polar line</td>
<td>Greater than 50% of mass across polar line (a) or mass crosses axial renal midline (b) or mass is entirely between polar lines (c)</td>
</tr>
<tr>
<td>(L)ocation relative to polar lines</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(H)ilar</td>
<td>Assigned if tumor touches main renal artery or vein</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**REFERENCES**


EDITORIAL COMMENT

These authors report external validation of the nomogram constructed by Kutikov et al that predicts the risks of malignancy and high grade features in renal masses (reference 12 in article). They noted reasonable predictive accuracy in a cohort of 391 patients. The study highlights 2 timely topics in urological decision making, of which one is methodological and the other is clinical.

Nomograms likely provide the most accurate risk estimates of all available prediction tools. However, concern has been raised that there is insufficient nomogram validation in external patient cohorts, which is akin to naming a starting quarterback based on his performance in practice rather than in live game action. Internal validation simply does not establish that a model is ready for prime time.

Clinically the recognition that treatment related morbidity may exceed therapeutic benefits while the subsequent movement toward surveillance protocols represents a shift toward a more sophisticated medical decision making paradigm. The continued development, validation and implementation of predictive models that can accurately discriminate between patients with a high or low likelihood of disease progression, particularly at high risk for competing cause mortality, will help clinicians to philosophically “do no harm” and minimize the potential hazards of therapy that are incurred by patients.

Brandon K. Isariyawongse
Glickman Urological and Kidney Institute and

Michael W. Kattan
Quantitative Health Sciences
Cleveland Clinic
Cleveland, Ohio