BJU Int 2024; **133:** 690–698 doi:10.1111/bju.16276

BJUI
BJU International

Original Article

Fully Automated Versions of Clinically Validated Nephrometry Scores Demonstrate Superior Predictive Utility versus Human Scores

Andrew M. Wood[1], Nour Abdallah[1], Nicholas Heller[2], Tarik Benidir[1], Fabian Isensee[3], Resha Tejpaul[2], Chalairat Suk-ouichai[4], Caleb Curry[1], Alex You[5], Erick Remer[6], Samuel Haywood[1], Steven Campbell[1], Nikolaos Papanikolopoulos[2] and Christopher Weight[1]

[1]Glickman Urological and Kidney Institute, Cleveland, OH, [2]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA, [3]German Cancer Research Center (DKFZ) Heidelberg, University of Heidelberg, Heidelberg, Germany, [4]Siriraj Hospital, Mahidol University, Bangkok City, Thailand, [5]Case Western Reserve University, and [6]Department of Diagnostic Radiology, Imaging Institute Cleveland Clinic, Cleveland, OH, USA

A.M.W. and N.A. contributed equally to this work and have earned the right to place their name first within their CV/resume.

Objective

To automate the generation of three validated nephrometry scoring systems on preoperative computerised tomography (CT) scans by developing artificial intelligence (AI)-based image processing methods. Subsequently, we aimed to evaluate the ability of these scores to predict meaningful pathological and perioperative outcomes.

Patients and Methods

A total of 300 patients with preoperative CT with early arterial contrast phase were identified from a cohort of 544 consecutive patients undergoing surgical extirpation for suspected renal cancer. A deep neural network approach was used to automatically segment kidneys and tumours, and then geometric algorithms were used to measure the components of the concordance index (C-Index), Preoperative Aspects and Dimensions Used for an Anatomical classification of renal tumours (PADUA), and tumour contact surface area (CSA) nephrometry scores. Human scores were independently calculated by medical personnel blinded to the AI scores. AI and human score agreement was assessed using linear regression and predictive abilities for meaningful outcomes were assessed using logistic regression and receiver operating characteristic curve analyses.

Results

The median (interquartile range) age was 60 (51–68) years, and 40% were female. The median tumour size was 4.2 cm and 91.3% had malignant tumours. In all, 27% of the tumours were high stage, 37% high grade, and 63% of the patients underwent partial nephrectomy. There was significant agreement between human and AI scores on linear regression analyses (R ranged from 0.574 to 0.828, all $P < 0.001$). The AI-generated scores were equivalent or superior to human-generated scores for all examined outcomes including high-grade histology, high-stage tumour, indolent tumour, pathological tumour necrosis, and radical nephrectomy (vs partial nephrectomy) surgical approach.

Conclusions

Fully automated AI-generated C-Index, PADUA, and tumour CSA nephrometry scores are similar to human-generated scores and predict a wide variety of meaningful outcomes. Once validated, our results suggest that AI-generated nephrometry scores could be delivered automatically from a preoperative CT scan to a clinician and patient at the point of care to aid in decision making.

Keywords

kidney imaging, artificial intelligence, machine learning, renal mass, nephrometry score

Introduction

Starting with the design and publication of the R.E.N.A.L. (radius, exophytic/endophytic, nearness to collecting system, anterior/posterior location, location relative to polar lines) and PADUA (Preoperative Aspects and Dimensions Used for an Anatomical classification of renal tumours) systems in 2009, nephrometry scoring systems have become important tools for describing renal mass complexity [1,2]. Initially intended to enhance surgeon communication and allow researchers to effectively measure and account for surgical difficulty, multiple studies have since demonstrated that nephrometry scores can be used to predict important oncological outcomes including histological grade, pathological staging, and patient survival [3,4]. Despite clear value to both risk assessment and surgical decision-making, widespread clinical adoption has been limited by interobserver variability and required time investment by busy clinicians [5,6]. Any technology that allows for mitigation of these two barriers would allow for wider spread adoption of nephrometry scoring systems and provide substantial value to Urologists.

The application of deep learning (DL), a subfield of machine learning (ML), to healthcare-related problems promises added value for questions that involve high-dimensional data [7–9]. Within the field of Urological Oncology, many impressive applications of DL thus far have involved renal cancer. For instance, DL has demonstrated an ability to reliably differentiate renal tumour subtypes and predict functional postoperative outcomes [10–13]. We have previously described a novel DL process for fully automated semantic segmentation of kidneys and kidney tumours [13]. This fully automated process promises significant potential benefits to researchers looking to investigate radiomic features on large scales not conducive to manual input. However, the potential advantages of automation in kidney cancer imaging research go beyond computer-generated tumour segmentation. For example, as a part of the previously described work, our group assessed the ability of artificial intelligence (AI)-generated R.E.N.A.L. nephrometry scores to predict pathological outcomes in patients undergoing partial (PN) or radical nephrectomy (RN) for renal mass. We found that AI-generated scores were highly correlated with scores calculated by human experts and could deduce the presence of malignancy, grade, stage, and necrosis just as reliably as human experts [14].

Despite the importance of this initial experience, it involved only one of many well-established nephrometry scores. In order for fully automated/AI-generated nephrometry scores and risk assessment models to obtain widespread adoption, demonstration of generalisability across multiple unique models is important. It was therefore our objective to develop, demonstrate, and validate the predictive utility of fully automated versions of multiple additional nephrometry score systems including the centrality index (C-Index), tumour contact surface area (CSA), and PADUA.

Patients and Methods

Cohort

Following ethics board approval of protocol code 1611M00821 at the University of Minnesota – Twin Cities, 544 consecutive adult patients undergoing extirpative surgery for a renal tumour at a single institution between 2010 and 2018 were identified. This cohort was used to host the previously reported KiTS19 international segmentation challenge. Overall inclusion and exclusion criteria for this cohort is based on a previously published KiTS19 (kidney and kidney tumour segmentation challenge) protocol [13]. Briefly, patients with a tumour thrombus were excluded (27 patients), as well as patients without an available arterial phase CT scan preoperatively (217 patients). Following the application of these exclusion criteria, a total of 300 patients remained. If a patient had more than one tumour removed, the largest tumour removed was used to determine nephrometry scores, as well as for pathological outcome determination. The full KiTS19 cohort, with scans, segmentations, clinical details, and outcomes, is now publicly available at https://kits-challenge.org/. Tumours of the 300 selected patients were subsequently assigned both AI-generated and human-calculated PADUA, C-Index, and tumour CSA scores based on the following process.

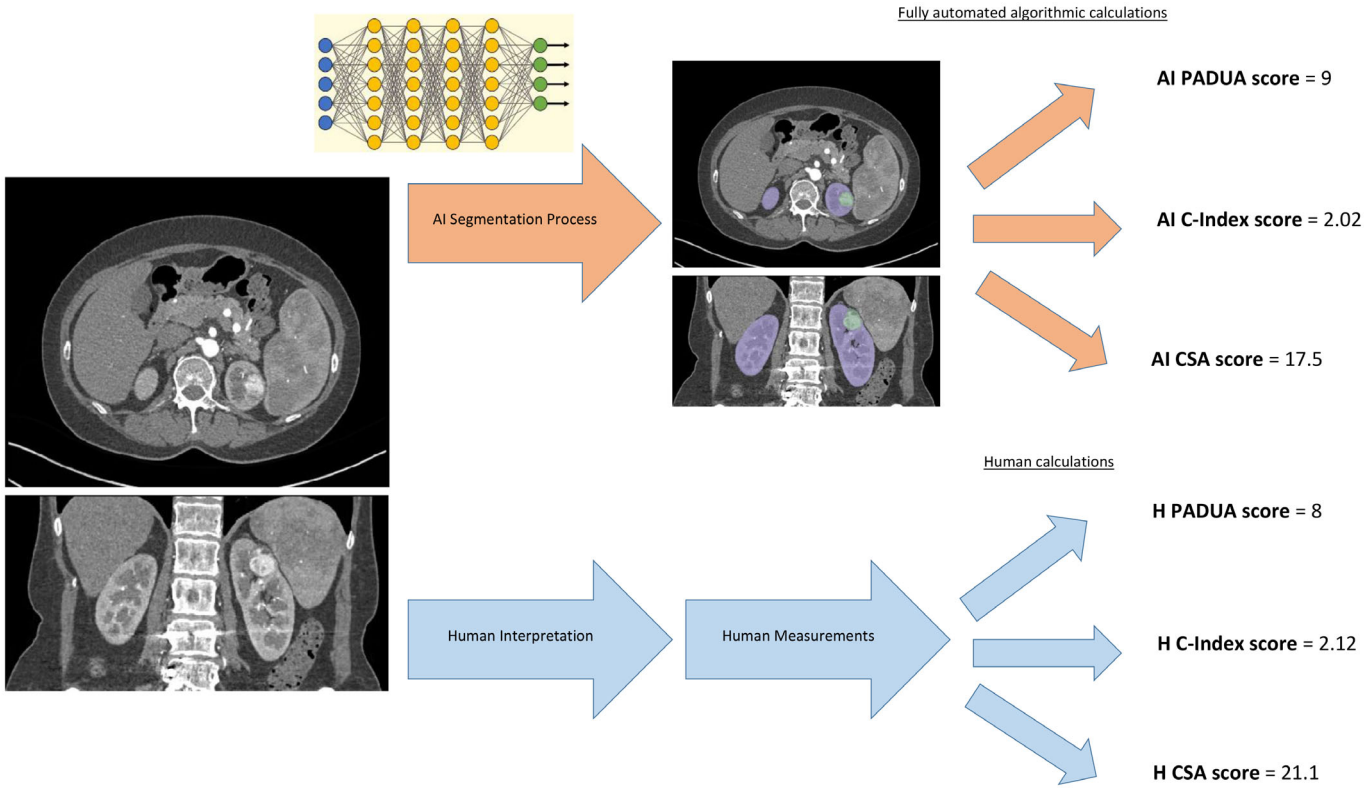
Artificial intelligence Scoring (AI Scores)

The following process was used to generate fully automated nephrometry scores from Digital Imaging and Communications in Medicine files without human intervention at any step. The process is summarised in Fig. 1.

Creating the 'ground truth' for DL and Selection of Fully Automated Segmentation Algorithm

Each slice of the arterial phase of included CT scans was manually annotated as 'kidney', 'kidney tumour', or 'background' by specialised medical personnel. The resulting dataset of almost 50 000 individual axial slices is referred to as the 'ground truth', providing the training material for the DL algorithms and the basis of extraction of nephrometry scores. These ground truth manual CT segmentations were then used to host an international segmentation challenge [13], with a single algorithm declared as the winner by demonstrating the highest level of fidelity to the ground truth segmentations [15]. Subsequently, this algorithm was used as the basis for generation of fully automated segmentations masks. As previously published, AI-generated segmentations

Fig. 1 Schematic representation demonstrating AI-generated segmentation if CT images, geometric algorithmic calculation of fully automated nephrometry scores, and corresponding human process.



provided segmentation masks similar to human-generated segmentations with a Sørensen-Dice coefficient of 0.92 [14]. An important potential source of bias within this process was the fact that the winning algorithm was trained on 210 of the 300 CT scans used in our study, introducing the possibility of biased predictions. Fortunately, since the winning team's method was based on an ensemble of five models constructed with five-fold cross-validation, we were able to use the individual models of this ensemble to obtain unbiased segmentation predictions for each of the 210 cases in the training cohort. From these masks, we extracted the components for nephrometry scores for the 300 patients described above. A more detailed methodology of the segmentation process was previously published [14].

Collecting Nephrometry Score Components from Segmented CT Scans

Preoperative Aspects and Dimensions Used for an Anatomical classification The six components of the PADUA score were generated from the fully automated segmentation masks using the following processes.

1. Location relative to sinus lines: renal sinus fat was defined as central voxels within the kidney segmentation with attenuation of low enough radiodensity to represent adipose tissue (≤−30.0 Hounsfield units [HU]). The top-most and bottom-most axial slices of the segmentation mask that contain at least one such voxel are identified and used as the sinus lines. The fraction of tumour voxels found on these slices or the slices between them is then returned. If >50% of tumour voxels were found between these lines, a score of 2 was assigned. Otherwise, a score of 1 was assigned.
2. Location relative to medial or lateral rim: the set of voxels in the kidney region that are adjacent to the 'background' region (perinephric fat) were determined and defined as the 'rim'. Throughout this paper, 'adjacent' voxels are defined as lying within each other's 8-neighbourhood (i.e., touching either side to side or at the corners) in the axial plane. The third dimension is not considered because variations in slice thickness could cause unintended artefacts. All pairs of rim and tumour voxels were compared to find the pair that are closest together, which represents the rim point nearest to the tumour. The position of this rim point was then compared with both the left–right position of the centroid of the affected kidney and the relative position of the overall left right midpoint of the axial slice. If the closest rim point was

located between the centroid of the kidney and the overall midline, the tumour was designated 'medial', and a score of 2 was assigned, otherwise it was given a designation of 'lateral' and a score of 1 was assigned.
3. Sinus involvement: the voxels in the aforementioned sinus region were considered and compared to voxels in the tumour region. If any sinus region voxels were located adjacent (defined as above) to one or more tumour region voxels, the tumour was said to be involving the renal sinus and a score of 2 was assigned. Otherwise, the tumour was said to not be involving the renal sinus and a score of 1 was assigned.
4. Collecting system involvement: the urinary collecting system (UCS) is generally understood to be bordered by sinus fat. We therefore defined the UCS as voxels within the predetermined area of sinus fat with attenuation exceeding the previously mentioned HU threshold (i.e., significantly brighter than the sinus fat). Similarly to renal sinus involvement, if UCS voxels were found adjacent (defined as above) to tumour voxels, the tumour was said to be involving the UCS and a score of 2 was assigned, otherwise a score of 1 was assigned.
5. Endophycity: for every axial slice that contained tumour voxels, the convex hull of the kidney region excluding the tumour was identified. For every tumour voxel, it was then recorded whether it lied inside (endophytic) or outside (exophytic) of this convex hull. Endophytic proportions were then quantised, with scores of 1, 2 or 3 assigned for endophytic proportions of <50%, 50–100%, and 100%, respectively.
6. Size: first, connected component analysis was used to identify the largest contiguous region of voxels labelled 'tumour' prior to the calculation of size, so any false-positive tumour voxels lying elsewhere in the image are not considered. Next, we identified the two tumour region voxels that were furthest apart and recorded this distance. Traditional PADUA cut-offs were then used to assign a score of 1, 2, or 3 for distances of <4, 4–7, and >7 cm, respectively.

Concordance Index The C-Index is calculated as the ratio of the distance between the centre of the mass and the centre of the kidney (centrality) to the radius of the tumour as a whole. The two components of this formula are generated from the automated segmentation masks using the following processes.

1. Centrality: centroids are calculated for the tumour region and the affected kidney region. Here, the affected kidney region is the union of the voxels labelled kidney and those labelled tumour. The distance between those centroids is calculated and recorded as the centrality value.
2. Radius: the distance between every pair of voxels in the tumour region is calculated using a simple vector norm. The maximum is returned. This value is then divided by two to obtain the radius value

Tumour contact surface area First, we found all voxels on the tumour region that border at least one voxel on the kidney region. The surface area of the resulting thin 'surface' region is determined by first transforming that region to a mesh object, and then simply summing the areas of the set of triangles that comprise that mesh. The result is divided by two to account for the fact that the thin surface object has both a front and back surface. This computation was performed using the implementation provided by the 'PyRadiomics' package [16].

Human Scoring

Traditional human-calculated nephrometry scores were assigned to each tumour by a team of three trained medical personnel blinded to the AI score results. A single score was generated for each nephrometry score on each CT scan. All personnel were trained in the three different nephrometry scoring systems by a fellowship trained Urological Oncologist (C.J.W.) prior to assigning nephrometry scores. All values were recorded as originally described and currently implemented in clinical practice [2,17,18].

Statistical Analysis

Simple linear regression analyses were performed to assess correlation between AI and human scores for tumour CSA and C-Index, and Pearson correlation coefficients (R) were calculated. For the ordinal PADUA score, Spearman's rank correlation coefficient was utilised to evaluate correlation between AI and human scores and reported as ρ (rho). Receiver operating characteristic (ROC) curves were developed from univariate logistic regression models to evaluate the discriminatory ability of AI and human scores to determine each outcome of interest. All statistical tests were performed on samples with N far greater than 30 observations, thus the central limit theorem was applied, and parametric statistical tests were deemed appropriate despite non-normality of sample data. Outcomes of interest included high stage (pathological [p]T Stage 3–4), high grade (Fuhrman Grade 3–4), indolent tumour (benign, or low stage and low grade), pathological tumour necrosis, surgical procedure, and high-grade complication as measured by the Clavien–Dindo classification. Notably, only patients undergoing PN were included in analyses of the nephrometry scores' ability to predict high-grade complications, as this is the context of the original development of nephrometry scoring systems. Areas under the ROC curves (AUCs) with CIs were calculated for both AI and human scores and plotted for comparison. R version 4.2.1 (R Foundation for Statistical Computing, Vienna, Austria), using the 'R

commander' package and 'R commander ROC' plug-in, was used for all statistical analyses.

Results

Characteristics of the Cohort

The baseline characteristics of the KiTS19 cohort are shown in Table 1. Of 300 patients, the median (interquartile range [IQR]) age was 60 (51–68) years, and 240 (60%) were males. The median (IQR) body mass index (BMI) was 29 (26–35) kg/m^2. In all, 188 patients (63%) underwent PN while the remaining 112 (37%) underwent RN. A laparoscopic or robotic approach was utilised in 221 (73%) of surgeries. Of the removed tumours, 275 (92%) were malignant; 75 (27%) were high stage (pT3–T4), 92 (37%) were high grade (Fuhrman Grade 3–4), and 69 (23%) contained pathological tumour necrosis. The median (IQR) diameter was 4.2 (2.6–6.1) cm. The mean (SD) estimated GFR change at ≥3 months after surgery was −13 (15.3) mL/min/1.73 m^2. Patients undergoing RN experienced a mean change of −24 mL/min/1.73 m^2 while those undergoing PN experienced a change of just −6 mL/min/1.73 m^2. The mean (SD) AI and human PADUA scores were 9.07 (1.58) and 9.13 (1.83), respectively. The mean (SD) AI and human C-Index scores were 1.86 (1.42) and 2.11 (1.38), respectively. The mean (SD) AI and human tumour CSA scores were 30.77 (26.04) and 54.01 (68.20), respectively.

Table 1 Baseline characteristics (N = 300).

Characteristic	Value
Gender, *n* (%)	
Female	120 (40)
Male	179 (60)
Transgender (male to female)	1 (0.3)
Age, years, median (IQR)	60 (51–68)
Tumour diameter, cm, median (IQR)	4.20 (2.60–6.12)
BMI, kg/m^2, median (IQR)	29 (26–35)
Baseline estimated GFR, mL/min/1.73 m^2, median (IQR)	72 (60–81)
AI C-Index score, mean (SD)	1.86 (1.42)
Human C-Index score, mean (SD)	2.11 (1.38)
AI tumour CSA score, mean (SD)	30.77 (26.04)
Human tumour CSA score, mean (SD)	54.01 (68.20)
AI PADUA score, mean (SD)	9.07 (1.58)
Human PADUA score, mean (SD)	9.13 (1.83)
Malignant renal mass, *n* (%)	275 (92)
Pathological T stage, *n* (%)	
0	24 (8.0)
1a	121 (40)
1b	60 (20)
2a	15 (5.0)
2b	5 (1.7)
3	8 (2.7)
3a	62 (21)
4	5 (1.7)
Tumour necrosis, *n* (%)	69 (23)
Tumour grade, *n* (%)	
0	25 (9.3)
1	33 (12)
2	119 (44)
3	66 (25)
4	26 (9.7)
Surgical technique, *n* (%)	
Laparoscopic	49 (16)
Open	79 (26)
Robotic	172 (57)
Nephrectomy type, *n* (%)	
PN	188 (63)
RN	112 (37)
Estimated blood loss, mL, median (IQR)	200 (100–400)
Length of hospital stay, days, median (IQR)	3.00 (2.00–4.00)
Postoperative complication by Clavien–Dindo Grade, *n* (%)	
Any	92 (31)
I	45 (15)
II	24 (8.0)
III	15 (5.0)
IV	5 (1.7)
V	2 (0.7)

Correlation of AI and Human Scores

There was strong correlation between AI and human scores for all three examined nephrometry scores. The AI PADUA was significantly associated with the human PADUA (ρ = 0.571, P < 0.001), the AI C-Index was significantly associated with the human C-Index (R = 0.828, P < 0.001), and the AI tumour CSA was significantly associated with the human tumour CSA (R = 0.757, P < 0.001). Least squares lines based on linear regressions for all three nephrometry scores can be found in Fig. 2. Within the PADUA score components, size was the most strongly correlated between the human and AI scores (ρ = 0.87, indicating strong correlation), while collecting system involvement was the weakest correlation (ρ = 0.245, indicating strong correlation). Details can be found in Table S1.

Comparison of Predictive Utility of AI and Human Scores

Table 2 summarises the discriminatory ability of the AI and human versions of the three nephrometry scores in predicting the previously mentioned pathological and surgical outcomes. Figure 3 shows the ROC curves for each score and outcome besides high-grade complication. The AI C-Index demonstrated a greater AUC than the human C-Index in predicting high-grade tumour, high-stage tumour, indolent tumour, pathological tumour necrosis, and surgical approach. The AI tumour CSA similarly demonstrated a greater AUC than the human tumour CSA in predicting high-grade tumour, high-stage tumour, indolent tumour, pathological tumour necrosis, and surgical approach. The AI PADUA demonstrated a greater AUC than the human PADUA in predicting pathological tumour necrosis and similar AUCs in predicting all other examined outcomes.

The subset of patients undergoing PN (188 in total) were separately examined to determine the ability of AI- and

Fig. 2 Correlation of AI and human nephrometry scores with least squares regression lines.

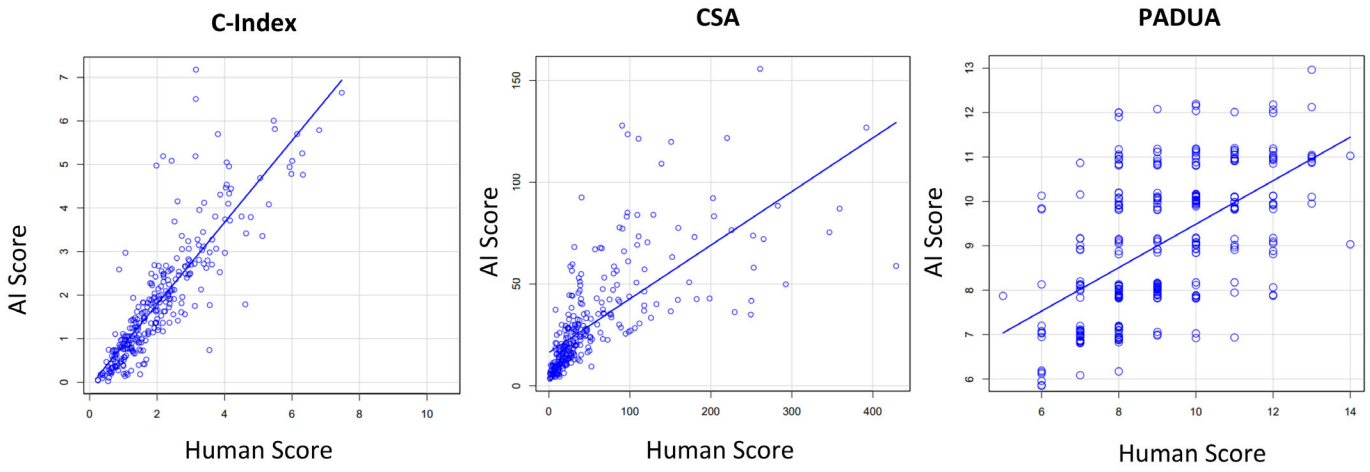


Table 2 Nephrometry scores predicting clinical outcomes

	AI AUC (95% CI)	*P*	Human AUC (95% CI)	*P*
C-Index				
High stage (pT3–4)	0.79	<0.001	0.76	<0.001
High grade (Fuhrman Grade 3 or 4)	0.75	<0.001	0.71	<0.001
Indolent (benign, or low grade and low stage)	0.76	<0.001	0.73	<0.001
Tumour necrosis	0.8	<0.001	0.76	<0.001
Surgical approach (RN vs PN)	0.87	<0.001	0.85	<0.001
High-grade complication (PN only)	0.54 (0.34–0.74)	0.722	0.58 (0.40–0.75)	0.491
Tumour CSA				
High stage	0.76	<0.001	0.73	<0.001
High grade	0.71	<0.001	0.69	<0.001
Indolent	0.74	<0.001	0.69	<0.001
Tumour necrosis	0.81	<0.001	0.75	<0.001
Surgical approach	0.86	<0.001	0.83	<0.001
High-grade complication (PN only)	0.50 (0.33–0.68)	0.508	0.52 (0.35–0.69)	0.494
PADUA score				
High stage	0.65	<0.001	0.71	<0.001
High grade	0.63	<0.001	0.64	<0.001
Indolent	0.64	<0.001	0.67	<0.001
Tumour necrosis	0.74	<0.001	0.71	<0.001
Surgical approach	0.71	<0.001	0.71	<0.001
High-grade complication (PN only)	0.54 (0.37–0.72)	0.660	0.54 (0.35–0.72)	0.531

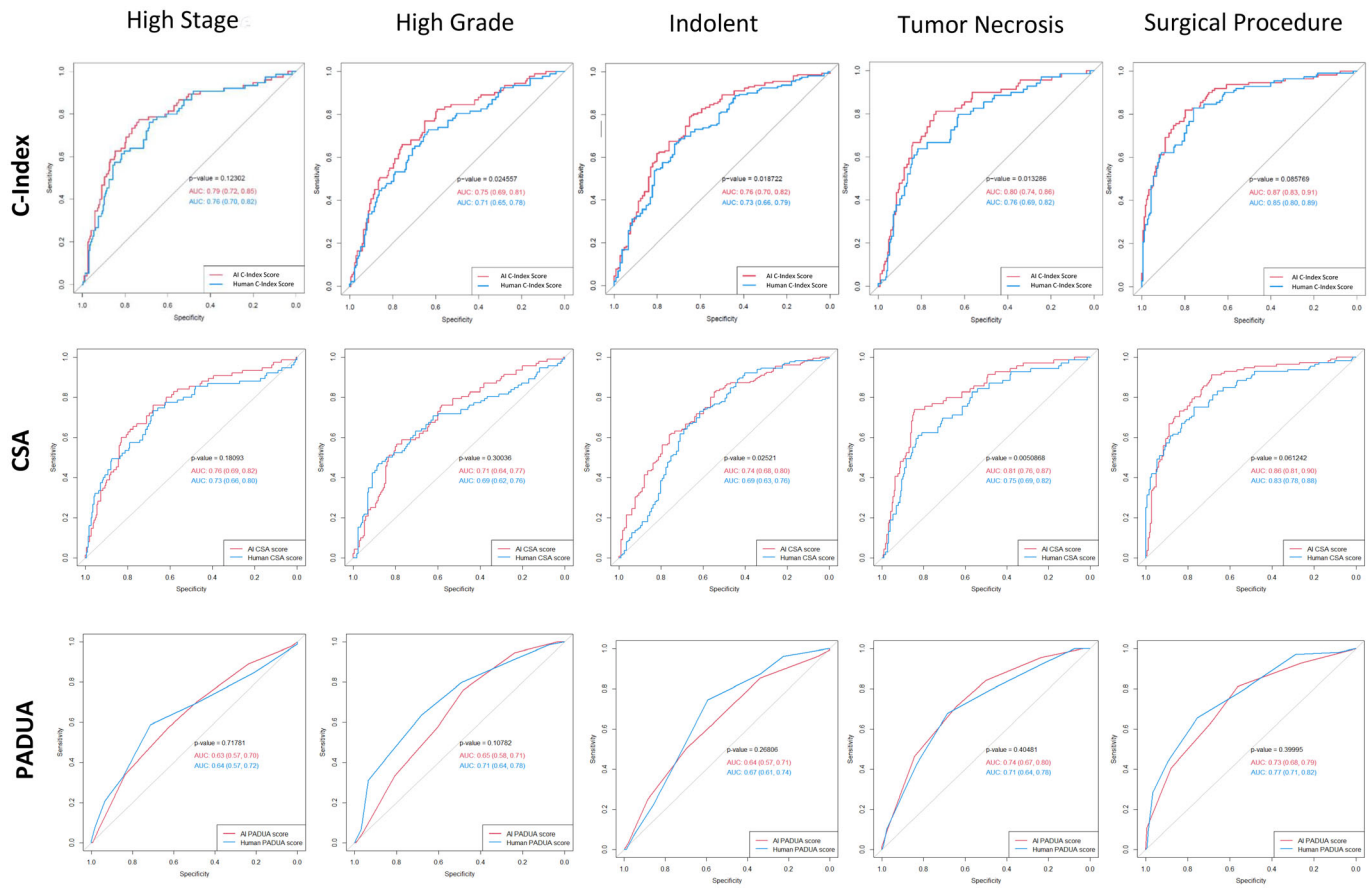
human-generated nephrometry scores to achieve the original intention of nephrometry scoring systems: predicting high-grade complications. The results of univariate logistic regression analyses on this subset for all six nephrometry scores can be found in Table 2. None of the examined nephrometry scores were significant predictors of high-grade complications following PN.

Discussion

Through the use of fully automated semantic segmentation followed by employment of three fully automated algorithms for calculation of different nephrometry scores, we were able to demonstrate a method for calculation of multiple unique nephrometry scores that required no human input. As previously reported, AI-generated segmentations provided segmentation masks that were impressively similar to human-generated segmentations (Sørensen-Dice coefficient = 0.92) [14]. Importantly, we were able to demonstrate that the PADUA, C-Index, and tumour CSA scores generated from these segmentations (also in a fully automated fashion) provide equivalent, and in many cases superior, predictive utility for clinical and pathological outcomes of interest.

Despite the relatively common inclusion of nephrometry scoring systems in renal mass research investigation, their clinical implementation has lagged behind. Unfortunately, despite efforts to categorise variables and simplify scoring systems, the largest impediment to widespread clinical use remains clinician time investment. Fortunately, the utilisation

Fig. 3 Receiver operating characteristic curves (with AUCs) for AI- vs human-generated nephrometry scores predicting pathological and perioperative outcomes.



of fully automated nephrometry scores promises to mitigate this issue while improving upon predictive utility. It is easy to envision a day where radiology reports for CT scans identifying a renal mass will include estimates of risk for pathological variables such as histological grade based on automated nephrometry score calculations that require minimal clinician or radiologist time investment. As automation and AI begin to play a larger role in diagnostic and prognostic risk assessment, it is vital that sufficient intermediary steps are taken to help clinicians trust predictions made by AI systems. We believe that a demonstration that a fully automated segmentation and score calculation process can generate prognostically useful scores for four completely different nephrometry scoring systems (R.E.N.A.L. was previously described) is an important step in this journey.

In addition to benefits to clinical implementation, automation of nephrometry scores carries important implications for continued renal mass and RCC research investigation. Widespread standardisation of renal mass complexity measurements within research efforts remains an important goal. Nephrometry scoring systems are the best tool available for this effort but are uncommonly deployed on very large-scale data sets because of the extensive time investment required. Fully automated nephrometry scores promise important improvements in efficiency and uniformity of kidney cancer research. With current methods, manually calculating nephrometry scores on a multi-thousand-patient renal mass dataset requires hundreds of person-hours. If a fully automated system, such as ours, is validated and popularised, a similar data set can be generated in 1–2 days. In addition to efficiency considerations, the use of an automated system immediately solves another inherent problem for nephrometry score-related research: interobserver variability in interpretation and measurement. As long as the algorithms used for segmentation and score calculation are the same, interobserver variability between systems is impossible.

Improvement in efficiency of calculation of imaging-related predictors of kidney cancer outcomes is particularly important in the current environment of AI-based investigation. The unique value of neural network approaches

to data examination is based on their ability to identify nuances between cases. This ability can only be realised with a sufficient sample size to capture how these nuances in variation relate to outcomes of interest [19]. While there is no single answer to the quantity of training data required for a given problem (i.e., predicting kidney cancer oncological outcomes), a higher volume of high-quality data generally produces superior results. Fully automated imaging evaluation of renal masses is vital for the development of research investigation in this space. In summary, the ability to generate multiple validated nephrometry scores in an automated fashion for large-volume data sets promises to improve the consistency, quality, and overall volume of renal mass complexity data for the field as a whole.

Our study is not without limitations. First, our algorithms could not generate all three nephrometry scores for nine of 300 cases (3%) due to an inability to find a lesion of interest. Whether these outliers were due to image quality issues, motion artefact, or other problems was not investigated. While it is important to recognise the technological limitations of an automated system, we believe a 3% failure rate is acceptable at this stage. Additionally, the AI algorithm used for segmentation was trained on a portion of the 300 involved CT scans, introducing the possibility of bias. However, use of the single algorithm (of five that make up the wider model) that was not exposed to each training set CT during five-fold cross validation reduces this bias. In addition, neither nephrometry scores nor clinical/pathological outcomes were involved in the training. Regardless, validation of these results on an independent data set is an important next step. Another potential limitation was our single-centre experience, as surgery was done at a single institution with >85% Caucasian patients and almost 50% harbouring clinical obesity (BMI >30 kg/m^2), which has important implications for generalisability. We evaluated for demographic (age, gender, BMI, race) effects on AI- to human-score differences and only found a weak ($R^2 = 0.058$, $P < 0.001$) relationship between BMI and PADUA score. While this provides some information that demographic factors are likely not playing a major role in our results, it is nevertheless important to consider for those who might have a significantly different patient population and is an important area for future study. Adding to generalisability, however, is the fact that CT scan images were collected in >70 medical facilities. We also utilised consecutive patients and thus accepted all tumour sizes, locations, and types, which reflects real-world practice. Finally, all segmentations and nephrometry score calculations were based on arterial phase CT scans. While this is not always available in real-world practice (often only nephrographic/venous phases are provided), using the arterial phase was necessary in order to best standardise the training set for the segmentation challenge. Further efforts on segmentation using venous and urographic phases are underway.

Conclusions

Fully automated kidney and tumour segmentation followed by geometric analysis is able to generate multiple nephrometry scores that are highly correlated with human-calculated scores. These scores were able to meet or exceed the predictive utility of human-calculated scores for clinically relevant outcomes. This has important implications for nephrometry score calculation efficiency, elimination of interobserver variability, and development of additional ML-based radiological investigation in the future. External validation of these fully automated results is necessary prior to full clinical implementation.

Disclosure of Interests

Nicholas Heller holds a position at Astrin Biosciences, Inc.

Funding

This work was supported in part by the National Cancer Institute of the National Institutes of Health under Award Number R01CA225435.

References

1 **Kutikov A, Uzzo RG.** The R.E.N.A.L. nephrometry score: a comprehensive standardized system for quantitating renal tumour size, location and depth. *J Urol* 2009; 182: 844–53

2 **Ficarra V, Novara G, Secco S et al.** Preoperative aspects and dimensions used for an anatomical (PADUA) classification of renal tumours in patients who are candidates for nephron-sparing surgery. *Eur Urol* 2009; 56: 786–93

3 **Weight CJ, Atwell TD, Fazzio RT et al.** A multidisciplinary evaluation of inter-reviewer agreement of the nephrometry score and the prediction of long-term outcomes. *J Urol* 2011; 186: 1223–8

4 **Kutikov A, Smaldone MC, Egleston BL et al.** Anatomic features of enhancing renal masses predict malignant and high-grade pathology: a preoperative nomogram using the RENAL nephrometry score. *Eur Urol* 2011; 60: 241–8

5 **Chapin BF, Wood CG.** The RENAL nephrometry nomogram: statistically significant, but is it clinically relevant? *Eur Urol* 2011; 60: 249–52

6 **Spaliviero M, Poon BY, Aras O et al.** Interobserver variability of R.E.N.A.L., PADUA, and centrality index nephrometry score systems. *World J Urol* 2015; 33: 853–8

7 **Rajkomar A, Dean J, Kohane I.** Machine learning in medicine. *N Engl J Med* 2019; 380: 1347–58

8 **Beam AL, Kohane IS.** Big data and machine learning in health care. *JAMA* 2018; 319: 1317–8

9 **LeCun Y, Bengio Y, Hinton G.** Deep learning. *Nature* 2015; 521: 436–44

10 **Kocak B, Yardimci AH, Bektas CT et al.** Textural differences between renal cell carcinoma subtypes: machine learning-based quantitative computed tomography texture analysis with independent external validation. *Eur J Radiol* 2018; 107: 149–57

11 **Feng Z, Rong P, Cao P et al.** Machine learning-based quantitative texture analysis of CT images of small renal masses: differentiation of angiomyolipoma without visible fat from renal cell carcinoma. *Eur Radiol* 2018; 28: 1625–33

12 **Sharma N, Zhang Z, Mir MC et al.** Comparison of 2 computed tomography-based methods to estimate preoperative and postoperative renal parenchymal volume and correlation with functional changes after partial nephrectomy. *Urology* 2015; 86: 80–6

13 **Heller N**, **Isensee F**, **Maier-Hein KH et al.** The state of the art in kidney and kidney tumour segmentation in contrast-enhanced CT imaging: results of the KiTS19 challenge. *Med Image Anal* 2021; 67: 101821
14 **Heller N**, **Tejpaul R**, **Isensee F et al.** Computer-generated R.E.N.A.L. nephrometry scores yield comparable predictive results to those of human-expert scores in predicting oncologic and perioperative outcomes [published correction appears in *J Urol.* 2022 Oct;208(4):939]. *J Urol* 2022; 207: 1105–15
15 **Isensee F**, **Jaeger PF**, **Kohl SAA**, **Petersen J**, **Maier-Hein KH**. nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* 2021; 18: 203–11
16 **Griethuysen JJM**, **Fedorov A**, **Parmar C et al.** Computational radiomics system to decode the radiographic phenotype. *Cancer Res* 2017; 77: e104–7
17 **Simmons MN**, **Ching CB**, **Samplaski MK**, **Park CH**, **Gill IS**. Kidney tumour location measurement using the C index method. *J Urol* 2010; 183: 1708–13
18 **Leslie S**, **Gill IS**, **de Castro Abreu AL et al.** Renal tumour contact surface area: a novel parameter for predicting complexity and outcomes of partial nephrectomy. *Eur Urol* 2014; 66: 884–93
19 **Rasmussen R**, **Sanford T**, **Parwani AV**, **Pedrosa I**. Artificial intelligence in kidney cancer. *Am Soc Clin Oncol Educ Book* 2022; 42: 1–11

Correspondence: Andrew M. Wood, Glickman Urological and Kidney Institute, 9500 Euclid Avenue, Cleveland, OH 44195, USA.
e-mail: wooda10@ccf.org

Abbreviations: AI, artificial intelligence; AUC, area under the ROC curve; BMI, body mass index; C-Index, concordance index; CSA, contact surface area; DL, deep learning; HU, Hounsfield units; IQR, interquartile range; ML, machine learning; (P)(R)N, (partial) (radical) nephrectomy; PADUA, Preoperative Aspects and Dimensions Used for an Anatomical classification; pT, pathological T stage; R.E.N.A.L., radius, exophytic/endophytic, nearness to collecting system, anterior/posterior location, location relative to polar lines; ROC, receiver operating characteristic; UCS, urinary collecting system.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Correlation of individual PADUA components between human and AI scores.

