European Radiology (2022) 32:4116–4127
https://doi.org/10.1007/s00330-021-08449-w

UROGENITAL

Malignancy risk stratification of cystic renal lesions based on a contrast-enhanced CT-based machine learning model and a clinical decision algorithm

Jérémy Dana[1,2,3,4] · Thierry L. Lefebvre[5,6] · Peter Savadjiev[4,5,7,8,9] · Sylvain Bodard[1,10] · Simon Gauvin[4,11] · Sahir Rai Bhatnagar[4,12] · Reza Forghani[4,9,11] · Olivier Hélénon[1,10] · Caroline Reinhold[4,9,11]

Received: 3 August 2021 / Revised: 17 October 2021 / Accepted: 29 October 2021 / Published online: 23 January 2022

Abstract

Objective To distinguish benign from malignant cystic renal lesions (CRL) using a contrast-enhanced CT-based radiomics model and a clinical decision algorithm.

Methods This dual-center retrospective study included patients over 18 years old with CRL between 2005 and 2018. The reference standard was histopathology or 4-year imaging follow-up. Training and testing datasets were acquired from two institutions. Quantitative 3D radiomics analyses were performed on nephrographic phase CT images. Ten-fold cross-validated LASSO regression was applied to the training dataset to identify the most discriminative features. A logistic regression model was trained to classify malignancy and tested on the independent dataset. Reported metrics included areas under the receiver operating characteristic curves (AUC) and balanced accuracy. Decision curve analysis for stratifying patients for surgery was performed in the testing dataset. A decision algorithm was built by combining consensus radiological readings of Bosniak categories and radiomics-based risks.

Results A total of 149 CRL (139 patients; 65 years [56–72]) were included in the training dataset—35 *Bosniak(B)-IIF* (8.6% malignancy), 23 *B-III* (43.5%), and 23 *B-IV* (87.0%)—and 50 CRL (46 patients; 61 years [51–68]) in the testing dataset—12 *B-IIF* (8.3%), 10 *B-III* (60.0%), and 9 *B-IV* (100%). The machine learning model achieved high diagnostic performance in predicting malignancy in the testing dataset (AUC = 0.96; balanced accuracy = 94%). There was a net benefit across threshold probabilities in using the clinical decision algorithm over management guidelines based on Bosniak categories.

Conclusion CT-based radiomics modeling accurately distinguished benign from malignant CRL, outperforming the Bosniak classification. The decision algorithm best stratified lesions for surgery and active surveillance.

Jérémy Dana and Thierry L. Lefebvre contributed equally

✉ Caroline Reinhold
caroline.reinhold@mcgill.ca

1 Assistance Publique - Hôpitaux de Paris, Paris University, Paris, France

2 Inserm U1110, Institut de Recherche Sur Les Maladies Virales Et Hépatiques, Strasbourg University, Strasbourg, France

3 Institute of Image-Guided Surgery, University Hospital Institute, Strasbourg, France

4 Present Address: Department of Diagnostic Radiology, McGill University, Montreal, Canada

5 Medical Physics Unit, McGill University, Montreal, Canada

6 Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK

7 School of Computer Science, McGill University, Montreal, Canada

8 Department of Pathology, McGill University, Montreal, Canada

9 Augmented Intelligence & Precision Health Laboratory of the Research Institute of McGill University Health Centre, Montreal, Canada

10 Department of Adult Radiology, Necker Enfant-Malades University Hospital, Assistance Publique Des Hôpitaux de Paris, Paris, France

11 Montreal Imaging Experts Inc, Montreal, Canada

12 Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada

Key Points

- *The radiomics model achieved excellent diagnostic performance in identifying malignant cystic renal lesions in an independent testing dataset (AUC = 0.96).*
- *The machine learning–enhanced decision algorithm outperformed the management guidelines based on the Bosniak classification for stratifying patients to surgical ablation or active surveillance.*

Keywords Machine learning · Kidney neoplasms · Kidney diseases cystic · Algorithms · Tomography X-ray computed

Abbreviations

3D	Three dimensional
AUC	Area under the receiver operating characteristic curve
CE-CT	Contrast-enhanced computed tomography
CI	Confidence interval
CRL	Cystic renal lesion
CT	Computed tomography
IBSI	Image Biomarker Standardization Initiative
ICC	Intra-class correlation coefficients
LASSO	Least absolute shrinkage and selection operator
MRI	Magnetic resonance imaging
RCC	Renal cell carcinoma
RQS	Radiomics quality score
VOI	Volume of interest

Introduction

Renal cysts have become a frequent finding on cross-sectional imaging due to the increased use of imaging in the work-up of suspected abdominal pathology and their high prevalence in elderly patients, approximately 50% of patients over 50 years of age [1]. Management and long-term imaging follow-up of renal cysts are an important burden on the healthcare system in terms of cost and utilization of imaging resources [2]. In addition, unnecessary surgical procedures performed on cysts that are ultimately proven to be benign may result in decreased renal function or morbidity [3–5]. Although most renal cysts are benign, up to 20% of complex cystic lesions may harbor a malignancy [6]. Conversely, complex cysts may result from complications such as hemorrhage or infection.

The standard radiological criterion for risk stratification of cystic renal lesions, known as the Bosniak classification system, was introduced in 1986 in an attempt to standardize the description of complex renal cysts and to provide classification guidelines for distinguishing nonsurgical from surgical cystic lesions. A fundamental limitation of the Bosniak classification system is the suboptimal grading of lesion complexity by visual assessment of cyst morphology including septal and wall thickness and irregularity. For instance, 55.1% of 887 Bosniak III lesions were found to be malignant in a recent meta-analysis [7], underscoring the uncertainty in interpretation of the Bosniak III category which often includes benign lesions such as epithelial or fibrous cysts, cystic nephromas, and oncocytomas, all of which may share imaging features with malignant lesions based on the Bosniak classification system. Thus, identifying the most predictive imaging features of malignancy remains essential to improve the diagnostic performance of the Bosniak classification system and reduce unnecessary surgery [8]. In 2019, an update of the Bosniak criteria was introduced with more discriminative definitions and novel quantitative criteria with the goals to: (1) improve the specificity of higher risk categories in order to increase the proportion of masses that are followed rather than resected, or ignored rather than followed, and (2) to provide specific definitions for individual terms to improve inter-reader agreement and promote cross-study consistency [9]. Preliminary studies comparing versions 2005 and 2019 have not yet validated this proposal [10–13]. Indeed, inter-reader variability and diagnostic performances were not markedly increased. Instead, a significant number of 2005 Bosniak III lesions, including malignancies, were reclassified as 2019 Bosniak IIF resulting in decreased sensitivity, increased specificity, and similar accuracy.

To overcome some of the limitations of qualitative visual image interpretation, quantitative image analysis methods, known as radiomics, have gained increased importance in radiology in recent years [14, 15]. Radiomics analyses are based on mathematically defined quantitative descriptors, not typically part of the radiologists' lexicon, which are computed automatically by image analysis algorithms. Only a few preliminary radiomics studies on renal cyst assessment based on diffusion-weighted MRI [16] and CE-CT [17] have been recently reported. These studies, however, have several important limitations including the absence of an independent testing dataset.

In this study, we hypothesized that quantitative image analysis using state-of-the-art radiomics extracted from CE-CT images of complex cystic renal lesions (CRL) could accurately and reproducibly predict malignancy. Therefore, our aim was to retrospectively build and determine the diagnostic performance of a radiomics-based logistic regression

model for the differentiation of benign from malignant complex renal cysts. The secondary objective was to propose a clinical decision algorithm combining the Bosniak classification system with the radiomics model.

Material and methods

The research project and access to retrospective health records without individual consent were approved by both research ethics boards (Institution 1, McGill University Health Centre, Research Ethics Board n°2020–5797; and Institution 2, Necker-Enfants Malades Hospital n°20191124194334).

Study participants and radiological assessment

This retrospective study was international and bi-institutional providing two independent datasets for training and testing. Patients with a renal cyst larger than at least 1 cm with no history of surgery or conditions associated with multiple renal cysts (e.g., autosomal dominant polycystic disease, Von Hippel-Lindau) were included. More than one cystic lesion could be included for each patient and lesions were then treated as independent. To ensure the generation of an optimal machine learning model applicable to daily clinical practice, the inclusion of patients was balanced to the Bosniak categories allowing us to provide the system with the full variability of imaging appearance while preventing fatal class imbalance between benign and malignant cases. Considering that the number of Bosniak III and IV lesions was the limiting factor given their lower prevalence, all patients with Bosniak III and IV CRL were first included to estimate the number of cases to be included for the remaining Bosniak categories. The estimated number of Bosniak I, II, and IIF cystic lesions was included in reverse-chronological order in a consecutive fashion. The testing dataset included approximately one-third of the number of cases included in the training dataset. Dedicated renal CT scans, including unenhanced scanning and enhanced imaging of corticomedullary (40–45 s), nephrographic (100–120 s), and excretory (approximately 8 min) phases, were extracted from each institution's radiological picture archiving communication system (IntelePACS, Intelerad Medical Systems Inc. and Vue PACS, Carestream Health Inc.). Qualitative visual assessment of renal cysts was performed by two trained abdominal radiologists blinded to pathology results and imaging follow-up in each dataset. Readers assigned Bosniak categories according to the classification system published in 2005 [18]. In case of disagreement, consensus was obtained between the two readers. The reference standard was the histopathological analysis of surgical specimens or follow-up over 4 years, extended to 5 years for multilocular CRL, by renal dedicated CE-CT or MRI without classification upgrade. All Bosniak III and IV lesions of the testing dataset had a histological diagnosis as the reference standard.

Quantitative methods

The quantitative radiomics analysis pipeline, trained and validated on the Institution 1 dataset, consisted of two main steps: (i) CRL segmentation and (ii) radiomics-based machine learning analysis, consisting of feature computation and selection, model building, and classification. In the development of the pipeline for 3D radiomics feature extraction and classification, we followed the recommendations from the Image Biomarker Standardization Initiative (IBSI) [19]. First, renal cysts were segmented semi-automatically on de-identified CE-CT images (nephrographic phase) using a commercial research software (Myrian®, Intrasense) with an implemented algorithm developed at Institution 1. After initial automatic segmentations, volumes of interest (VOIs) were manually corrected around the gross cystic volume by a junior radiologist (J.D.) with a 3-year experience in interpreting dedicated renal CT scans. An assessment of radiomics feature reproducibility was conducted. Copies of the original image data were created and subjected to variation in imaging parameters, i.e., changes in voxel size resampling and image gray-level discretization. Furthermore, VOIs were subjected to size changes (i.e., erosion and dilation) to assess the robustness of radiomics features to contour changes. Radiomics features were extracted with the PyRadiomics package [20] for each combination of image preprocessing parameters, and intra-class correlation coefficients (ICC) [21] were then used to determine the set of image preprocessing parameters leading to the highest features' reproducibility and robustness as a function of VOI changes. The image analyst performing radiomics feature extraction was blinded to pathology results and imaging follow-up in each dataset.

For the subsequent analysis, a single choice of imaging parameters (voxel size and image gray-level discretization) leading to features with the highest ICC values was retained, and only radiomics features with ICC > 0.8 were included. Finally, from these selected features, only the most discriminating ones were kept for the final model, as determined with a least absolute shrinkage and selection operator (LASSO) regression. LASSO feature selection was employed for its ability to reduce the dimension of high-dimensional radiomics data while selecting few uncorrelated features [22]. Stratified ten-fold cross-validated training of a L1-regularized logistic regression model was performed on balanced samples of the training dataset using the selected radiomics features.

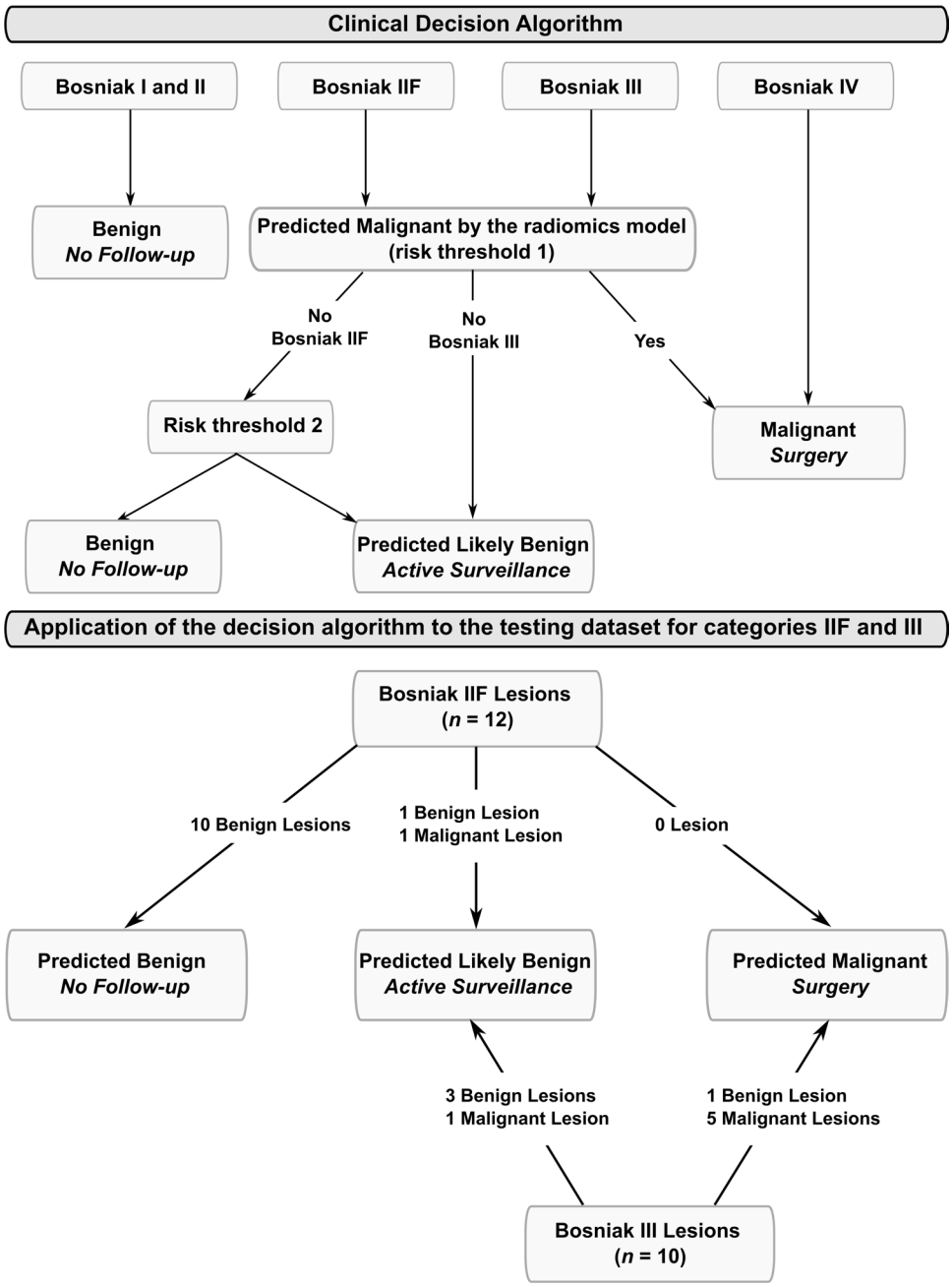

Fig. 1 Clinical decision algorithm applied on the testing dataset (Bosniak categories IIF and III). The algorithm starts by taking as an input the 2005 Bosniak category determined by the radiologist. Lesions categorized as Bosniak I and II are deemed benign and assessed as not requiring any follow-up. The radiomics-based risk threshold maximizing Youden's index for the differentiation of benign from malignant cysts in the training dataset is employed for prescribing surgical ablation to Bosniak IIF and III lesions exceeding this risk threshold (risk threshold 1). Another radiomics-based risk threshold defined by the average risk plus a standard deviation of the radiomics risk calculated for Bosniak I and II lesions in the training dataset is used to identify benign Bosniak IIF lesions requiring no follow-up from those "likely" benign but requiring active surveillance (risk threshold 2). Finally, Bosniak IV are all considered malignant by the decision algorithm and assessed as requiring surgery. When applied to the testing dataset, the clinical decision algorithm improved the management of Bosniak IIF and III lesions

This final model was then tested on an independent testing dataset (Institution 2) to assess its diagnostic performance for distinguishing benign from malignant CRL. Diagnostic performance metrics, reported in the training and testing datasets, included areas under the receiver operating characteristic (AUC), sensitivity, specificity, balanced accuracy, positive predictive value, and negative predictive value for the threshold maximizing Youden's index to maximize balanced discrimination of cysts and prevent both the under-diagnosis of malignant lesions and unnecessary surgeries of benign lesions [23]. Loess calibration curves for binary outcomes were plotted with 95% bootstrapped confidence intervals to assess the agreement between the radiomics-predicted risk of malignancy and the reference standard–confirmed malignancy status in the training and testing datasets. Calibration was either mean, weak, moderate, or strong [24]. The integrated calibration index (ICI) for binary logistic regression models was evaluated as the difference between observed and predicted risks weighted by the probability density function of predicted risks [25].

All logistic regression modeling was conducted in Python 3.7.4 using the Scikit-learn machine learning package [26]. The radiomics quality score (RQS) of this study was assessed based on published criteria [27].

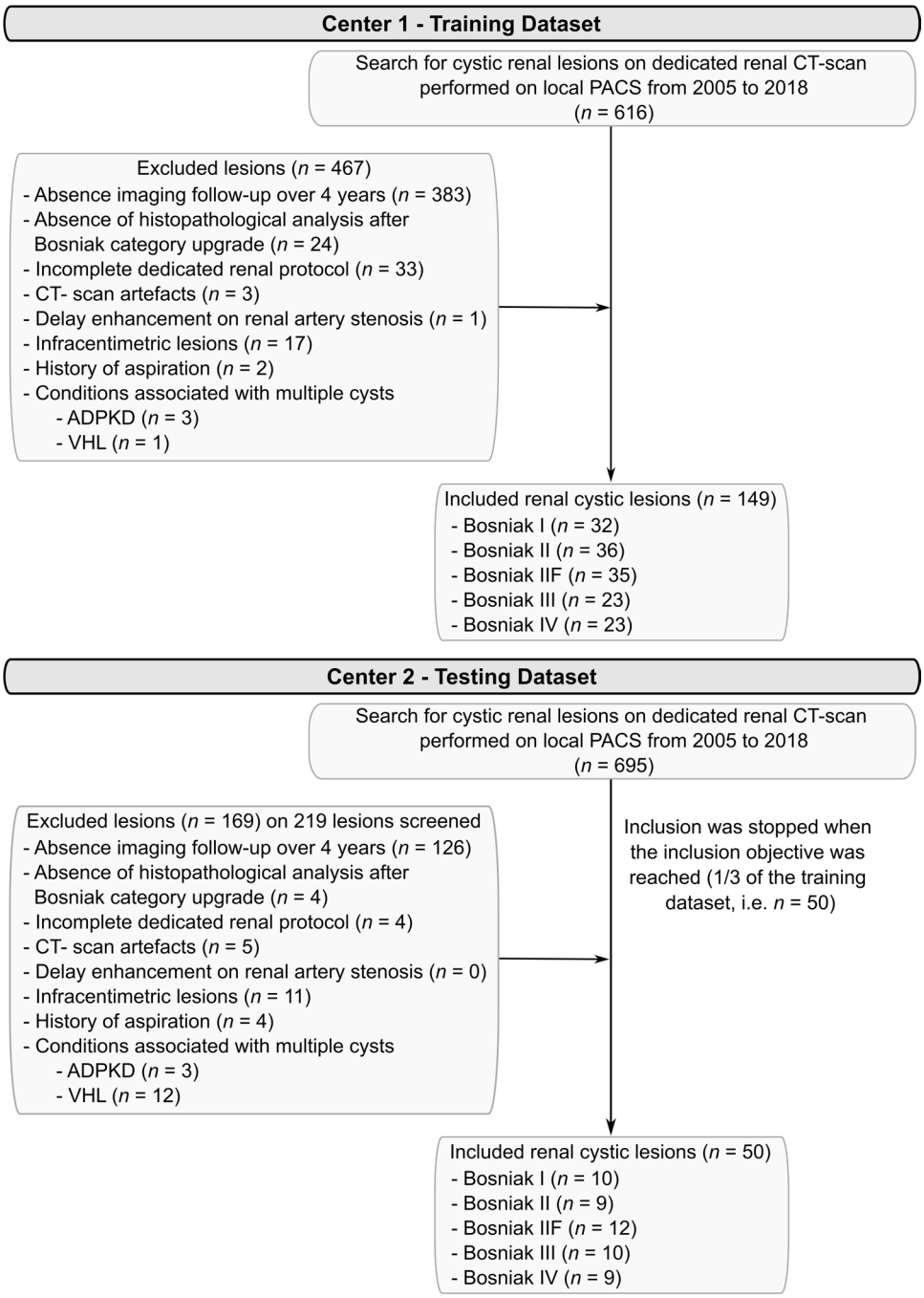

Fig. 2 Flowchart of patient selection. Renal cysts at each center were included consecutively in reverse-chronological order for categories I, II, and IIF considering the number of included Bosniak III and IV lesions to balance the number of cysts in each category. ADPKD, autosomal dominant polycystic kidney disease; VHL, Von Hippel-Lindau)

Radiomics and clinical decision algorithm

To assist physicians in better stratifying patients for serial imaging follow-up or ablative therapies, a clinical decision algorithm was built on the training dataset (Fig. 1).

Decision curve analysis was then conducted to determine the clinical usefulness of the radiomics model, the management guidelines based on the Bosniak classification, and the decision algorithm combining both in the stratification of Bosniak IIF and III lesions for surgery or active surveillance [28].

Clinical impact of the decision algorithm was assessed in the external testing dataset.

Full details of the materials and methods are provided in the online supplement.

Table 1 Patient characteristics distributed according to Bosniak categories. Pathology of renal cysts that underwent surgery is detailed. All lesions categorized as Bosniak I or II were benign. Every upgraded lesion during imaging follow-up had histopathological analysis. All Bosniak III and IV lesions of the testing dataset had histopathological confirmation. *RCC*, renal cell carcinoma; *IQR*, interquartile range. *Three of the 23 Bosniak IV lesions were benign (metanephric adenoma, angiomyolipoma with epithelial cysts, oncocytoma)

	Bosniak I		Bosniak II		Bosniak IIF		Bosniak III		Bosniak IV	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing
Renal cysts (*n*)	32	10	36	9	35	12	23	10	23	9
Patients (*n*)	27	9	34	9	33	10	23	9	22	9
With ipsilateral cysts	2	0	0	0	2	1	0	0	0	0
With contralateral cysts	3	1	2	0	0	0	0	1	1	0
Age (median, IQR)	66 [55–73]	64 [58–67]	66 [56–71]	61 [50–63]	65 [60–70]	61 [56–69]	59 [51–70]	45 [40–55]	68 [62–74]	68 [62–72]
Male (*n*, %)	20 (74)	7 (78)	25 (74)	8 (89)	20 (61)	7 (70)	14 (61)	4 (44)	14 (64)	6 (67)
Reference standard										
4-year follow-up (*n*, %)	32 (100)	10 (100)	36 (100)	9 (100)	30 (86)	10 (83)	10 (43)	0	0	0
Median follow-up in years (IQR)	8 (6–10)	6 (5–8)	8 (5–9)	5 (5–7)	5 (4–8)	6 (5–7)	7 (5–10)	-	-	-
Histopathological analysis (*n*, %)	0	0	0	0	5 (14)	2 (17)	13 (57)	10 (100)	23 (100)	9 (100)
Mean interval with 1st dedicated CT in months (range)					34 (0–73)	14 (6–21)	6 (0–22)	5 (0–12)	4 (0–20)	2 (0–7)
Malignancy and histopathology-confirmed tumor type										
Malignant (*n*, %)	0	0	0	0	3 (8.6)	1 (8)	10 (43)	6 (60)	20 (87)*	9 (100)
Clear cell RCC	-	-	-	-	2 (66)	-	7 (70)	3 (50)	15 (75)	5 (56)
Papillary RCC	-	-	-	-	1 (33)	-	3 (30)	-	4 (20)	4 (44)
Eosinophilic RCC	-	-	-	-	-	-	-	1 (17)	-	-
Unclassified RCC	-	-	-	-	-	-	-	-	1 (5)	-
Multilocular cystic renal neoplasm of low malignant potential	-	-	-	-	-	1 (100)	-	2 (33)	-	-

Results

Population

From 2005 to 2018, we included 149 CRL in the training set and 50 CRL in the testing set (Fig. 2; Table 1). In the training dataset, the median age was 65 years (inter-quartile range, 56–72 years) and 67% ($n=93$) of patients were men. In the independent testing dataset, the median age was 61 years (inter-quartile range, 51–68 years) and 67% ($n=31$) of patients were men. Twenty-two percent ($n=33$) of the lesions were malignant in the training dataset, and

Table 2 Estimates of diagnostic performance of the logistic regression model using 4 radiomics features and the Bosniak classification system for distinguishing benign from malignant renal cysts with inter-reader agreement. 95% bootstrapped confidence intervals are in parentheses

		AUC	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)
Radiomics model	Training	0.96 (0.91–0.98)	100 (95–100)	88 (81–93)	93 (86–97)	70 (60–83)	100 (98–100)
	Testing	0.96 (0.89–1.00)	88 (65–100)	97 (84–100)	94 (84–98)	93 (72–100)	94 (80–100)
Bosniak classification 2005	Training	–	90 (77–98)	86 (78–92)	87 (80–92)	65 (51–78)	97 (91–100)
	Testing	–	93 (61–100)	87 (73–97)	88 (78–96)	72 (55–94)	97 (81–100)

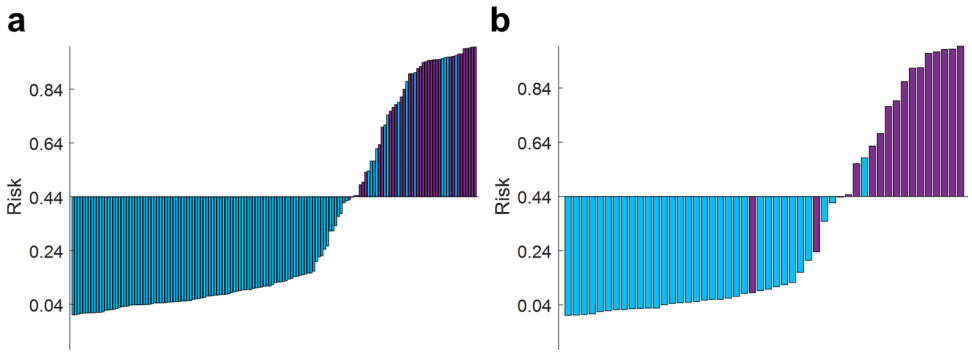

Fig. 3 Bar charts of radiomics-based risks in the (**a**) training and (**b**) testing datasets (blue bars are benign cases and purple bars are malignant cases)

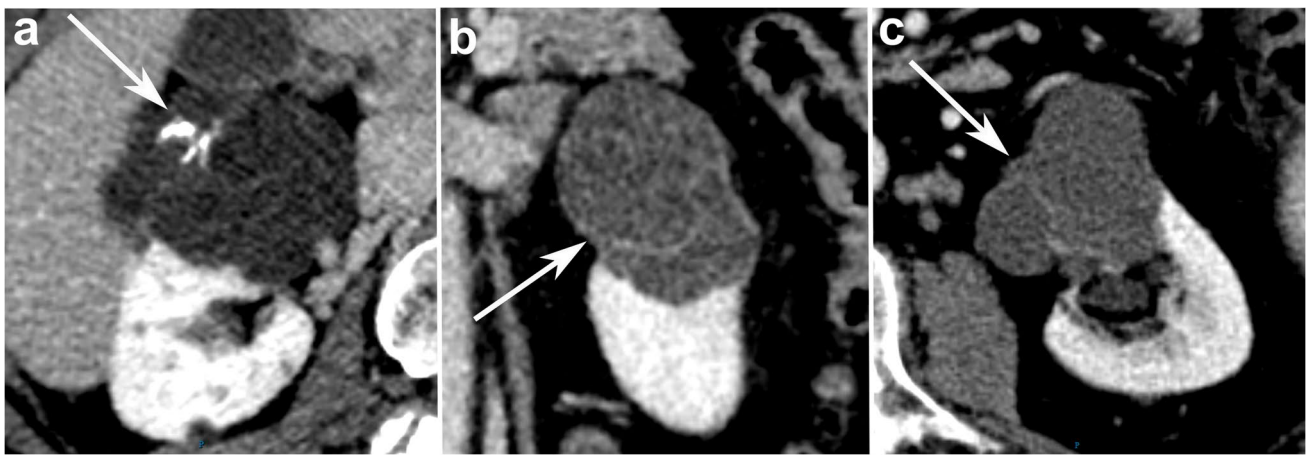

Fig. 4 Bosniak IIF complex cystic renal lesions (testing dataset) on nephrographic phase contrast-enhanced computed tomography assessed with radiological Bosniak classification version 2005, and the radiomics model. **a** Benign cystic renal lesion correctly predicted benign by the radiomics model with a radiomics risk of 7.2% and not requiring follow-up. Axial contrast-enhanced CT scan (nephrographic phase) demonstrates a few thin septa (no measurable enhancement) with thick calcifications (arrow). **b** Benign cystic renal lesion correctly predicted benign by the radiomics model with a radiomics risk of 34.8% but classified as requiring imaging follow-up by the decision algorithm. Coronal contrast-enhanced CT scan (nephrographic phase) demonstrates a few thin septa (arrow) with no measurable enhancement. **c** Multilocular cystic renal neoplasm of low malignant potential wrongly predicted benign by the radiomics model but with an increased malignancy risk (23.6%) compared to the average risk of all benign 2005 Bosniak IIF lesions (9.6% in the testing set). Axial contrast-enhanced CT scan (nephrographic phase) demonstrated a few thin septa with no measurable enhancement (arrow)

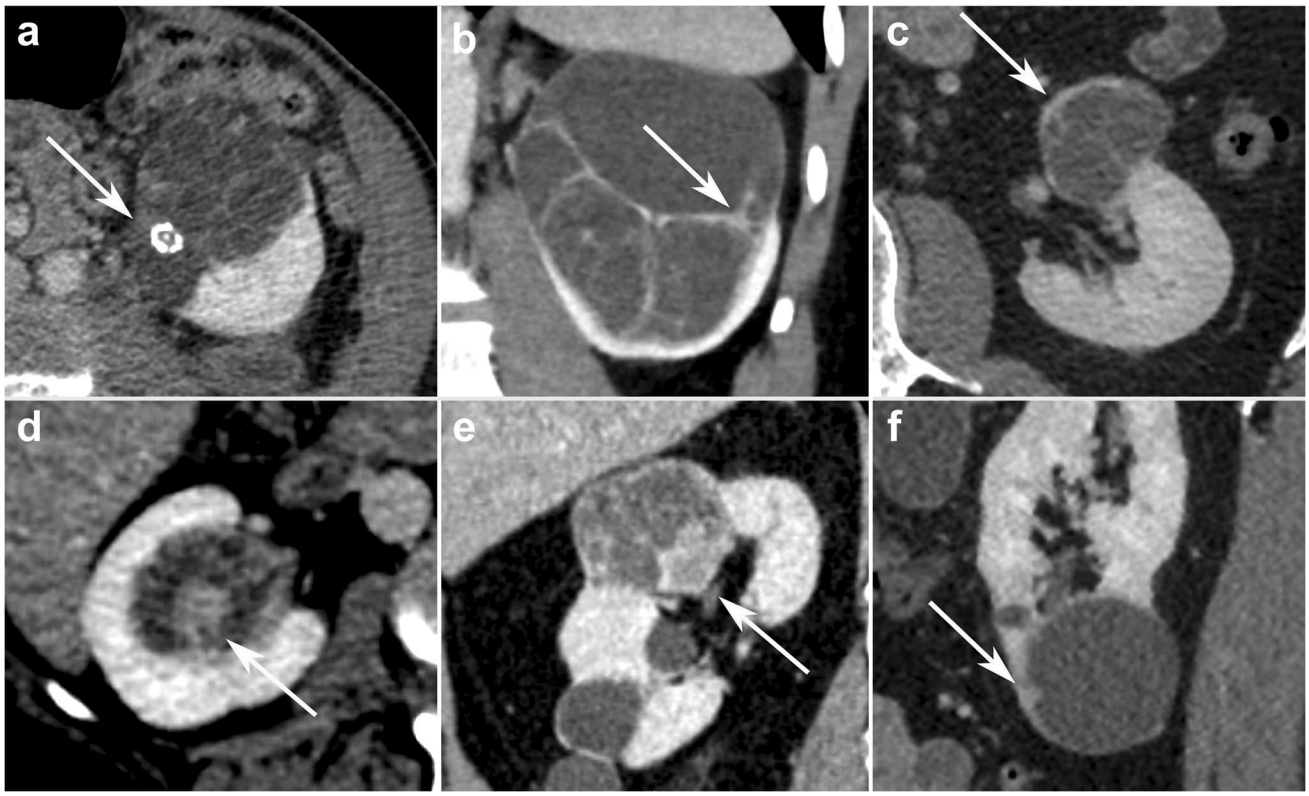

Fig. 5 Bosniak III and IV complex cystic renal lesions (testing dataset) on nephrographic phase contrast-enhanced computed tomography assessed with radiological Bosniak classification version 2005, and the radiomics model. **a** Cystic nephroma, classified as 2005 Bosniak III, correctly predicted as benign with a radiomics risk of 9.3%. Axial contrast-enhanced CT scan (nephrographic phase) demonstrated a multilocular cystic mass with multiple smooth minimally thickened enhancing septa and a coarse calcification (arrow). **b** Clear cell renal cell carcinoma Fuhrman II/IV, classified as 2005 Bosniak III, correctly predicted malignant with a radiomics risk of 79.3%. Coronal contrast-enhanced CT scan (nephrographic phase) demonstrates multiple thickened irregular enhancing septa (arrow). **c** Clear cell renal cell carcinoma Fuhrman II/IV, classified as 2005 Bosniak III, correctly predicted malignant with a radiomics risk of 86.4%. Axial contrast-enhanced CT scan (nephrographic phase) demonstrates multiple thickened irregular enhancing septa and wall (arrow). **d** Clear cell renal cell carcinoma Fuhrman III/IV, classified as 2005 Bosniak IV, correctly predicted malignant with a radiomics risk of 99.5%. Axial contrast-enhanced CT scan (nephrographic phase) demonstrates multiple thickened enhancing septa and wall with a nodular soft tissue component (arrow). **e** Clear cell renal cell carcinoma Fuhrman III/IV, classified as 2005 Bosniak IV, correctly predicted malignant with a radiomics risk of 97.3%. Coronal contrast-enhanced CT scan (nephrographic phase) demonstrates multiple minimally thickened irregular enhancing septa and a nodular soft tissue component (arrow). **f** Clear cell renal cell carcinoma Fuhrman II/IV, classified as 2005 Bosniak IV, correctly predicted with a radiomics risk of 98%. Axial contrast-enhanced CT scan (nephrographic phase) demonstrated a nodular thickening of the wall (arrow)

32% ($n = 16$) in the testing dataset. The median lesion size was 32 mm (inter-quartile range, 19–55 mm) in the training dataset and 43 mm (23–66) in the testing dataset.

Bosniak classification

The Bosniak classification system achieved good diagnostic performance in the training cohort with a sensitivity of 90% and a specificity of 86%; and also in the testing cohort with a sensitivity of 93% and a specificity of 87% (Table 2). Malignancy rates in each Bosniak category are summarized in Table 1.

Radiomics analysis

Ten-fold cross-validated LASSO feature selection led to a final model consisting of the 4 most discriminative radiomics features. These radiomics features included 2 first-order histogram statistics, the 10th percentile and the range of voxel intensities; 1 shape feature, surface-volume ratio; and 1 texture feature, dependence entropy from gray-level dependence matrices (GLDM).

Training of the regularized logistic regressor resulted in high diagnostic performance for predicting malignancy of CRL. The training of the logistic regressor resulted in the following model expressing the risk of malignancy as:

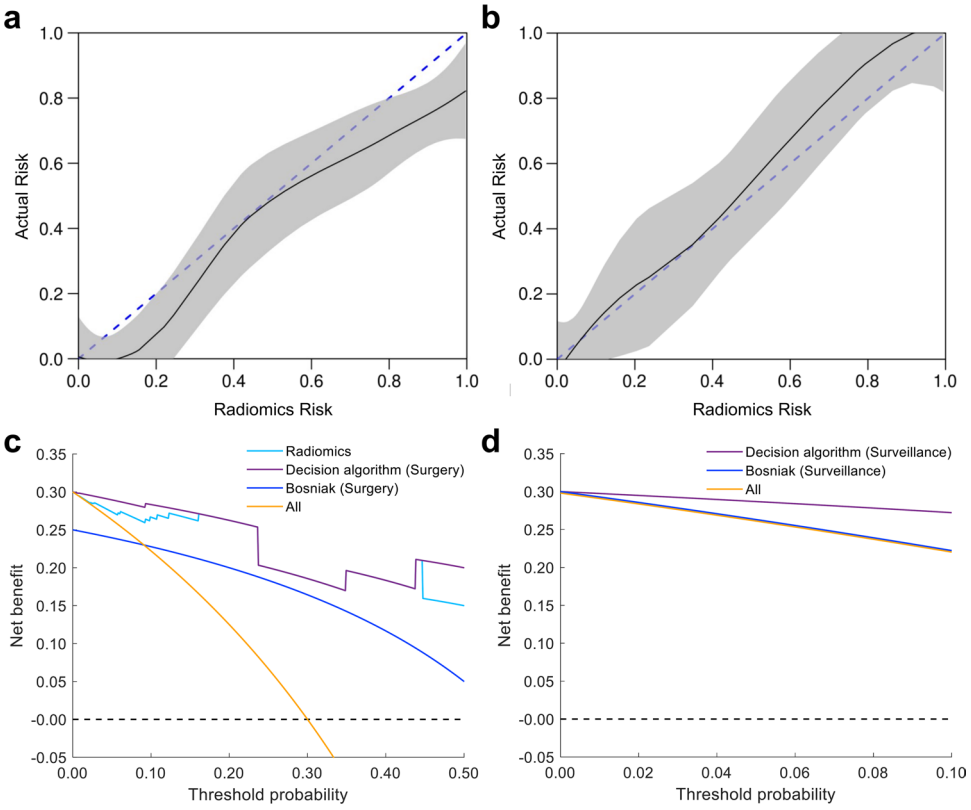

Fig. 6 Calibration curves of radiomics-based machine learning modeling for identifying malignant complex cystic renal lesions on nephrographic phase contrast-enhanced computed tomography images in (**a**) the training dataset and (**b**) the testing dataset. Decision curves of radiomics modeling, management guidelines based on the Bosniak classification, decision algorithm, and treat-all scheme on the external testing dataset (**c**) for the prescription of surgery and (**d**) for the indication of active surveillance in Bosniak IIF and III lesions. In calibration curves (**a**, **b**), gray areas represent 95% bootstrapped confidence intervals and dashed blue lines represent the ideal calibration. In the decision curve analysis, the net benefit was calculated as the proportion of true positive malignant lesions subject to undergoing surgery minus the proportion of false positive lesions weighted by the relative harm of no surgery compared with the negative consequences of an unnecessary surgery $\frac{p_t}{1-p_t}$. In the decision curves, the dashed black line corresponds to the absence of net benefits of a treat-none scheme, shown for reference. The net benefit was quantified at different threshold probabilities in the testing dataset, up to 0.5 and 0.1 regarding (**c**) surgery and (**d**) active surveillance, respectively, as it would otherwise be unreasonable to forgo the indicated management

$\log(\frac{\text{risk}}{1-\text{risk}}) = -1.44 + 1.87 \times \text{DependenceEntropy} - 1.58 \times \text{SurfaceVolumeRatio} + 1.28 \times 10\text{Percentile} - 0.86 \times \text{Range}$.

When applied to the external testing dataset, this classifier performed robustly (Table 2). Radiomics-based risks in the training and testing datasets are shown in bar charts in Fig. 3.

Based on the final radiomics-based machine learning model applied to the testing dataset, misclassifications were only observed in Bosniak categories IIF (1 malignant lesion; Fig. 4) and III (1 benign and 1 malignant lesion; Fig. 5). Calibration curves of the radiomics model estimating the risk of malignancy demonstrated moderate calibration in both training and testing datasets, associated with flexible risk prediction and success in avoiding overfitting. They showed overall good agreement with a small departure from the ideal fit in the training (ICI = 0.084) and testing cohorts (ICI = 0.033) (Fig. 6). The RQS of the radiomics analysis of this study was 18.

Clinical impact of decision algorithm

Applied to the testing dataset, the decision algorithm combining the discriminative power of radiomics risks with the current radiological Bosniak assessment (version 2005) provided higher net benefit across all threshold probabilities in terms of correctly stratifying patients to surgical ablation or active surveillance than the management guidelines based on the Bosniak classification only or treat-all scheme in Bosniak IIF and III lesions (Fig. 6). Figure 1 illustrates the impact of the decision algorithm on Bosniak IIF and III lesions. This is consistent with the higher specificity observed for radiomics modeling compared to that of Bosniak classification only, suggesting an improved decision support when using this algorithm combining both methods. One of the malignant

Bosniak III lesions was wrongly predicted as benign using only the radiomics model with a very low risk (0.08) but still assigned to active surveillance by the dual decision algorithm demonstrating its usefulness. Further details are reported in the online supplementary material.

Discussion

Although the Bosniak classification system correlates well with the risk of malignancy [29], it remains limited for grading lesion complexity within the IIF and III categories [7], resulting in unnecessary surgeries or serial follow-ups. In this retrospective dual-center study, the developed radiomics-based machine learning model achieved excellent diagnostic performance in distinguishing benign from malignant CRL, outperforming the Bosniak classification system. The inclusion of only 4 reproducible and discriminating radiomics features in a regularized logistic regression model resulted in robust and consistent performance between training and testing datasets from different institutions, with moderate calibration. If prospectively validated, the proposed decision algorithm, combining radiological readings of Bosniak categories with radiomics-based risk analysis, has the potential to decrease the current burden of CRL on the healthcare system by identifying low-risk Bosniak IIF lesions requiring no follow-up and redefining them as Bosniak II lesions while high-risk Bosniak III lesions would be subject to surgical resection. Bosniak IIF and III lesions of intermediate risk would be subject to active surveillance. We propose an annual follow-up over a 4-year period as time to progression of Bosniak IIF lesions was demonstrated to range from 6 months to 3.2 years (average 19.2 months) by Hindman et al. [30]. Active surveillance is considered a reasonable option as cystic renal cell carcinomas grow slowly, rarely presenting with metastases [31–33]. As the Bosniak classification has been proven highly effective for category I, II, and IV lesions in both academic and community hospitals, the radiomics risk should not alter patient management in these categories, although the machine learning model accurately predicted the pathologic (benign versus malignant) outcome in all these categories. Nevertheless, equivalent diagnostic performance of the machine learning model compared to the Bosniak classification cannot be definitely established in this small cohort. Hence, the proposed algorithm represents a safe strategy and may prevent unnecessary invasive therapies in the clinical setting.

The robustness of the radiomics model developed in this study relies on many factors: (1) the reproducibility of radiomics features was assessed without any regards to their diagnostic capacities to remove selection bias and to improve the reliability of the final modeling; (2) this pipeline complied with recommendations from the IBSI; (3) a logistic regressor was chosen among all machine learning models to favor interpretability and simplicity over complexity of other methods prone to overfitting on the training dataset, thus facilitating the translation and the repeatability of the modeling; (4) the use of histopathology or long-term follow-up as the reference standard, which is the clinical standard-of-care for identifying malignancy of CRL, thus confirming the clinical relevance of the findings of this study; and (5) the use of an external independent testing dataset assured the translatability and robustness of the developed model as shown by the consistently high diagnostic performance seen on both the training and testing datasets. In fact, the radiomics quality score (RQS) of this study was high compared to the average RQS reported in a recent meta-analysis on radiomics in renal cancer CT/MR imaging (RQS, 18 vs. 3.4) [34]. It is also interesting to note that the final 4 selected radiomics features reflect quantitative visual patterns important in assessing malignancy in cystic renal lesions by trained radiologists.

The main study limitation was the imperfect reference standard. Indeed, although a 4-year follow-up is sufficient to conclude benignity in the majority of complex cystic renal lesions of low risk, some rare CRL of low malignant potential may present changes after 4-year follow-up. To reduce the risk of biases, we extended the follow-up to 5 years for multilocular CRL. It is also important to highlight that all Bosniak III and IV lesions of the testing dataset underwent pathologic examination. Infra-centimetric lesions were excluded as they are too small to be characterized on CT scan [35]. We acknowledge that the integration of the Bosniak classification system in the decision algorithm implies inter-reader variability. However, the potential benefits of combining the machine learning model with the radiological reading outweigh these limitations. The updated 2019 version of the Bosniak classification aims to address the issue of inter-reader variability and improve the diagnostic performance of predicting malignancy [9]. The proposed classification has yet to be validated and, eventually, uniformly implemented [10–13]. If validated, it could replace the 2005 classification in a next version of the decision algorithm. Another limitation lies in the retrospective collection of data and the relatively small sample size of the training dataset. Nevertheless, testing the model on an independent dataset demonstrated its strong diagnostic performance. Finally, assignment of Bosniak category of renal cysts in the training and testing datasets was performed by two different pairs of subspecialty-trained abdominal radiologists. However, rather than being a limitation of the study results, it reinforces the value and generalizability of the proposed radiomics model and decision algorithm given that despite the inter-reader variability of the Bosniak classification [6], the decision algorithm demonstrated strong and consistent performance across the training and testing datasets.

This decision algorithm must be prospectively validated in a multicenter study including Bosniak IIF and III lesions. This study demonstrates that combining visual assessment by trained radiologists based upon standard guidelines with a robust machine learning pipeline potentiates mutual strengths. The strength of our proposed decision algorithm relies on an optimized association of quantitative biomarkers and the subjective but valuable experience of radiologists.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00330-021-08449-w.

Funding This work was jointly funded by the Fonds de recherche du Québec—Santé (FRQS) and the Fondation de l'association des radiologistes du Québec (FARQ). Thierry L. Lefebvre has received support from the Natural Sciences and Engineering Research Council of Canada to conduct this work.

Declarations

Guarantor The scientific guarantor of this publication is Caroline Reinhold.

Conflict of interest The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry One of the authors has significant statistical expertise.

Informed consent Written informed consent was waived by the Institutional Review Board.

Ethical approval Institutional Review Board approval was obtained.

Methodology
- retrospective
- diagnostic or prognostic study
- multicenter study

References

1. Kissane JM (1976) The morphology of renal cystic disease. Perspect Nephrol Hypertens 4:31–63
2. Smith AD, Carson JD, Sirous R et al (2019) Active surveillance versus nephron-sparing surgery for a Bosniak IIF or III renal cyst: a cost-effectiveness analysis. AJR Am J Roentgenol 212:830–838. https://doi.org/10.2214/AJR.18.20415
3. Sun M, Bianchi M, Hansen J et al (2012) Chronic kidney disease after nephrectomy in patients with small renal masses: a retrospective observational analysis. Eur Urol 62:696–703. https://doi.org/10.1016/j.eururo.2012.03.051
4. Van Poppel H, Da Pozzo L, Albrecht W et al (2007) A prospective randomized EORTC intergroup phase 3 study comparing the complications of elective nephron-sparing surgery and radical nephrectomy for low-stage renal cell carcinoma. Eur Urol 51:1606–1615. https://doi.org/10.1016/j.eururo.2006.11.013
5. Tan H-J, Norton EC, Ye Z et al (2012) Long-term survival following partial vs radical nephrectomy among older patients with early-stage kidney cancer. JAMA 307:1629–1635. https://doi.org/10.1001/jama.2012.475
6. El-Mokadem I, Budak M, Pillai S et al (2014) Progression, interobserver agreement, and malignancy rate in complex renal cysts (≥Bosniak category IIF). Urol Oncol 32:24.e21-24.e27. https://doi.org/10.1016/j.urolonc.2012.08.018
7. Sevcenco S, Spick C, Helbich TH et al (2017) Malignancy rates and diagnostic performance of the Bosniak classification for the diagnosis of cystic renal lesions in computed tomography - a systematic review and meta-analysis. Eur Radiol 27:2239–2247. https://doi.org/10.1007/s00330-016-4631-9
8. Benjaminov O, Atri M, O'Malley M et al (2006) Enhancing component on CT to predict malignancy in cystic renal masses and interobserver agreement of different CT features. AJR Am J Roentgenol 186:665–672. https://doi.org/10.2214/AJR.04.0372
9. Silverman SG, Pedrosa I, Ellis JH et al (2019) Bosniak classification of cystic renal masses, version 2019: an update proposal and needs assessment. Radiology 292:475–488. https://doi.org/10.1148/radiol.2019182646
10. Park MY, Park KJ, Kim M-H, Kim JK (2021) Bosniak classification of cystic renal masses version 2019: comparison to version 2005 for class distribution, diagnostic performance, and inter-reader agreement using CT and MRI. AJR Am J Roentgenol. https://doi.org/10.2214/AJR.21.25796
11. Pacheco EO, Torres US, Alves AMA et al (2020) Bosniak classification of cystic renal masses version 2019 does not increase the interobserver agreement or the proportion of masses categorized into lower Bosniak classes for non-subspecialized readers on CT or MR. Eur J Radiol 131:109270. https://doi.org/10.1016/j.ejrad.2020.109270
12. Yan JH, Chan J, Osman H et al (2021) Bosniak classification version 2019: validation and comparison to original classification in pathologically confirmed cystic masses. Eur Radiol. https://doi.org/10.1007/s00330-021-08006-5
13. Tse JR, Shen J, Shen L et al (2020) Bosniak classification of cystic renal masses version 2019: comparison of categorization using CT and MRI. AJR Am J Roentgenol 216:412–420. https://doi.org/10.2214/AJR.20.23656
14. Savadjiev P, Chong J, Dohan A et al (2019) Image-based biomarkers for solid tumor quantification. Eur Radiol. https://doi.org/10.1007/s00330-019-06169-w
15. Dana J, Agnus V, Ouhmich F, Gallix B (2020) Multimodality imaging and artificial intelligence for tumor characterization: current status and future perspective. Semin Nucl Med. https://doi.org/10.1053/j.semnuclmed.2020.07.003
16. Gillingham N, Chandarana H, Kamath A et al (2019) Bosniak IIF and III renal cysts: can apparent diffusion coefficient-derived texture features discriminate between malignant and benign IIF and III cysts? J Comput Assist Tomogr 43:485–492. https://doi.org/10.1097/RCT.0000000000000851
17. Miskin N, Qin L, Matalon SA et al (2020) Stratification of cystic renal masses into benign and potentially malignant: applying machine learning to the Bosniak classification. Abdom Radiol (NY). https://doi.org/10.1007/s00261-020-02629-w
18. Israel GM, Bosniak MA (2005) An update of the Bosniak renal cyst classification system. Urology 66:484–488. https://doi.org/10.1016/j.urology.2005.04.003
19. Zwanenburg A, Leger S, Vallières M, Löck S (2020) Image biomarker standardisation initiative. Radiology 191145. https://doi.org/10.1148/radiol.2020191145
20. van Griethuysen JJM, Fedorov A, Parmar C et al (2017) Computational radiomics system to decode the radiographic phenotype. Cancer Res 77:e104–e107. https://doi.org/10.1158/0008-5472.CAN-17-0339
21. Bartko JJ (1966) The intraclass correlation coefficient as a measure of reliability. Psychol Rep 19:3–11

22. Tibshirani R (1996) Regression shrinkage and selection via the LASSO. J R Stat Soc Ser B Methodol 58:267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
23. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27:861–874. https://doi.org/10.1016/j.patrec.2005.10.010
24. Van Calster B, Nieboer D, Vergouwe Y et al (2016) A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol 74:167–176. https://doi.org/10.1016/j.jclinepi.2015.12.005
25. Austin PC, Steyerberg EW (2019) The integrated calibration index (ICI) and related metrics for quantifying the calibration of logistic regression models. Stat Med 38:4051–4065. https://doi.org/10.1002/sim.8281
26. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
27. Lambin P, Leijenaar RTH, Deist TM et al (2017) Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 14:749–762. https://doi.org/10.1038/nrclinonc.2017.141
28. Vickers AJ, Cronin AM, Elkin EB, Gonen M (2008) Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Med Inform Decis Mak 8:53. https://doi.org/10.1186/1472-6947-8-53
29. Reese AC, Johnson PT, Gorin MA et al (2014) Pathological characteristics and radiographic correlates of complex renal cysts. Urol Oncol 32:1010–1016. https://doi.org/10.1016/j.urolonc.2014.02.022
30. Hindman NM, Hecht EM, Bosniak MA (2014) Follow-up for Bosniak category 2F cystic renal lesions. Radiology 272:757–766. https://doi.org/10.1148/radiol.14122908
31. Mousessian PN, Yamauchi FI, Mussi TC, Baroni RH (2017) Malignancy rate, histologic grade, and progression of Bosniak category III and IV complex renal cystic lesions. AJR Am J Roentgenol 209:1285–1290. https://doi.org/10.2214/AJR.17.18142
32. Chandrasekar T, Ahmad AE, Fadaak K et al (2018) Natural history of complex renal cysts: clinical evidence supporting active surveillance. J Urol 199:633–640. https://doi.org/10.1016/j.juro.2017.09.078
33. Pruthi DK, Liu Q, Kirkpatrick IDC et al (2018) Long-term surveillance of complex cystic renal masses and heterogeneity of Bosniak 3 lesions. J Urol 200:1192–1199. https://doi.org/10.1016/j.juro.2018.07.063
34. Ursprung S, Beer L, Bruining A et al (2020) Radiomics of computed tomography and magnetic resonance imaging in renal cell carcinoma-a systematic review and meta-analysis. Eur Radiol 30:3558–3566. https://doi.org/10.1007/s00330-020-06666-3
35. Herts BR, Silverman SG, Hindman NM et al (2018) Management of the incidental renal mass on CT: a white paper of the ACR Incidental Findings Committee. J Am Coll Radiol 15:264–273. https://doi.org/10.1016/j.jacr.2017.04.028