Cancer Biomarkers -1 (2024) 1–11
DOI 10.3233/CBM-230360
IOS Press

Radiomics and artificial intelligence for risk stratification of pulmonary nodules: Ready for primetime?

Roger Y. Kim
Division of Pulmonary, Allergy, and Critical Care, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
Tel.: +1 215 662 3677; E-mail: roger.kim@pennmedicine.upenn.edu

Received 15 August 2023
Accepted 17 January 2024

Abstract. Pulmonary nodules are ubiquitously found on computed tomography (CT) imaging either incidentally or via lung cancer screening and require careful diagnostic evaluation and management to both diagnose malignancy when present and avoid unnecessary biopsy of benign lesions. To engage in this complex decision-making, clinicians must first risk stratify pulmonary nodules to determine what the best course of action should be. Recent developments in imaging technology, computer processing power, and artificial intelligence algorithms have yielded radiomics-based computer-aided diagnosis tools that use CT imaging data including features invisible to the naked human eye to predict pulmonary nodule malignancy risk and are designed to be used as a supplement to routine clinical risk assessment. These tools vary widely in their algorithm construction, internal and external validation populations, intended-use populations, and commercial availability. While several clinical validation studies have been published, robust clinical utility and clinical effectiveness data are not yet currently available. However, there is reason for optimism as ongoing and future studies aim to target this knowledge gap, in the hopes of improving the diagnostic process for patients with pulmonary nodules.

Keywords: Radiomics, artificial intelligence, lung cancer, risk stratification, pulmonary nodule

1. Introduction

The past four decades have seen a dramatic increase in thoracic computed tomography (CT) imaging, resulting in approximately 1.6 million adults in the U.S. diagnosed with incidentally-detected pulmonary nodules (PNs) annually [1,2]. Moreover, based on the 2013 U.S. Preventative Services Task Force (USPSTF) recommendations, an estimated 8.0 million adults in the U.S. are eligible for lung cancer screening with low-dose CT [3], and this population is anticipated to expand to 15.1 million with implementation of the 2021 USPSTF recommendations [4]. When PNs are detected – either incidentally or via screening – lung cancer is the primary concern, as it is the deadliest malignancy in the U.S. and worldwide [5]. A definitive diagnosis of lung cancer requires an invasive lung biopsy, which is associated with certain procedural costs and potential significant risks, including respiratory failure, pneumothorax, myocardial infarction, and even death [6,7,8,9,10]. Therefore, malignancy risk stratification is the fundamental first step in guiding PN management decisions among clinicians, who seek to diagnose cancer in a timely manner while avoiding unnecessary procedures for those with benign PNs [11]. Among suspicious PNs > 8 mm in maximal diameter, clinical guidelines for both incidentally-detected PNs (American College of Chest Physicians [12]. Fleischner Society [13]) and screen-detected PNs (American College of Radiology Lung Imaging Reporting and Data System [Lung-RADS] [14]) recommend surgical resection for high risk PNs ($> 65\%$ risk of malignancy) and conservative management with non-invasive serial CT imaging surveillance for very low risk PNs ($< 5\%$

risk of malignancy). However, clinicians face a diagnostic dilemma among intermediate risk (5%–65% malignancy risk) PNs, as they must decide whether to pursue a lung biopsy or surveil with serial imaging. This crucial decision has important implications for patients. A malignant PN inappropriately managed with imaging surveillance delays a cancer diagnosis and may even deny a patient the opportunity for curative treatment. On the other hand, a patient with a benign PN recommended to undergo lung biopsy has been unnecessarily exposed to the risks and costs associated with this invasive procedure.

There currently exists a significant misalignment between malignancy risk stratification processes and clinical management decisions [9,15,16,17,18]. As many as 45% of individuals undergoing a lung biopsy for evaluation of a PN are ultimately found to have a benign diagnosis [15,16,17,18,19,20], meaning that a considerable proportion of patients are unnecessarily exposed to the potential complications and harms of lung biopsy procedures. While conventional regression-based risk prediction models incorporating a variety of clinical and PN characteristics have been in existence since the late 1990s (e.g., the Mayo Clinic and Brock models) [21,22,23,24] they do not reliably outperform routine clinician assessment of malignancy risk [15,25,26]. Moreover, only 18% of thoracic surgeons and 31% of pulmonologists regularly use any clinical risk prediction model [27], and clinicians do not consistently document a quantitative estimate of cancer risk [28]. Thus, a core focus of the thoracic oncology scientific and clinical community is to improve PN malignancy risk stratification to better guide subsequent management decisions [29,30]. A smorgasbord of biomarkers has been developed recently [31] including blood-based [32,33,34,35,36], airway-based [19,37,38], breath-based [39,40,41], and imaging-based tests and devices [42,43]. This article will focus specifically on recent efforts to use radiomics and artificial intelligence (AI) technology for PN risk stratification and the practical hurdles that exist for clinical implementation.

2. Radiomics and artificial intelligence

Radiomics-based computer-aided diagnosis (CAD) tools demonstrate promise for noninvasive PN risk stratification using solely CT imaging data. CAD describes the automation of image review to assist clinicians with making diagnoses [44], and the past two decades have seen CAD paired with radiomics, which uses advanced mathematical analysis of imaging data to aid interpretation [45,46]. More recently, the evolution of AI has allowed deep learning using neural networks to enhance the development of radiomics-based CAD tools [47,48]. The potential benefit of such tools lies in their ability to analyze additional data invisible to the human eye (including shape, spatial complexity, textures, and wavelet transformations) and provide information to clinicians beyond PN size, spiculation, and density [49]. Additionally, in contrast to traditional clinical risk prediction models that require clinicians to enter discrete variables into a model to calculate a probability of malignancy, radiomics-based CAD tools automate this process, which theoretically could lower the threshold for clinical uptake. Numerous studies to date have been published describing the development and validation of radiomics-based biomarkers for PN risk stratification [50,51,52,53,54,55,56,57,58,59,60,61,62,63]. An exhaustive systematic review of all radiomics-based CAD tools is outside the scope of this focused narrative review, which will cover select notable examples to date (Table 1).

Initial efforts to incorporate radiomics-based quantitative imaging data into models to distinguish between malignant and benign PNs used conventional machine learning approaches, which rely on explicit parameters based on expert knowledge and classic multivariable model development techniques. In 2006, Way and colleagues first described a CAD system that was trained on clinical imaging data, evaluated using data from the Lung Image Database Consortium and differentiated malignant from benign PNs using morphological and texture characteristics via a three-dimensional active contour method, achieving an area under the receiver operating characteristic curve (AUC) of 0.83 [51]. This system was then updated a few years later to include additional nodule characteristics including surface smoothness and shape irregularity, achieving an AUC of 0.86 [52]. Next, this group performed a multi-reader, multi-case study using retrospective PN CT data from the University of Michigan to evaluate the effect of this CAD tool on radiologists' performance discriminating between malignant and benign PNs and found that on average radiologists' AUC increased from 0.83 to 0.85 ($P < 0.01$) [53]. In 2018, Huang and colleagues published the results of their CAD algorithm, which analyzed adjacent lung tissues in addition to PN texture features and was derived from random forest machine learning using National Lung Screening Trial (NLST) data [54]. They performed a matched case-control study and reported a CAD AUC of 0.92, sensitivity of 0.95,

Table 1
Selected studies on radiomics-based risk stratification of pulmonary nodules

Publication	Study design and objective	Populations or datasets	Model and analytical details	Key results
Way et al, 2006 [51]	Analytical validation study to develop a CAD model and assess performance of image segmentation.	Training data: 96 PNs (4–60 mm; 46% malignant) from 58 pts at the University of Michigan. Validation data for segmentation: experienced radiologists' segmentation of 23 PNs from LIDC.	3D active contour segmentation with manual feature extraction, selection, and classification. CAD model trained and tested using leave-one-case-out resampling scheme.	AUC = 0.83. Model-segmented PN volumes greater than those outlined by LIDC radiologists.
Way et al, 2009 [52]	Analytical validation study to refine above CAD model.	Training data: 256 PNs (3–38 mm; 48% malignant) from 152 pts at the University of Michigan.	Novel PN surface features characterizing smoothness and shape irregularity added to CAD model (described above). Demographics (age, gender) and LDA classifier also assessed.	AUC = 0.86 with addition of novel PN surface features. No significant difference in CAD model performance when demographic features or LDA classifier included.
Way et al, 2010 [53]	Retrospective multi-reader, multi-case study to assess effect of above CAD model on radiologists' performance discriminating between malignant and benign PNs.	Reader study: 6 fellowship-trained thoracic radiologists evaluated 256 PNs (3–38 mm; 48% malignant) from 152 pts at the University of Michigan.	CAD model (described above). Model output = relative malignancy rating on a scale of 1 to 10, representing a 10-bin histogram of scores with fitted Gaussian distributions for malignant benign PNs.	CAD AUC = 0.86. Average radiologists' AUC increased from 0.83 to 0.85 with CAD.
Huang et al, 2018 [54]	Analytical validation study using matched case-control data to derive and evaluate a novel CAD model.	Training data: 140 PNs (4–20 mm; 50% malignant) from 140 pts in the NLST. Validation data: 46 PNs (4–20 mm; 43% malignant) from 46 pts in the NLST. All pts underwent lung biopsy. Malignant and benign PNs were matched based on demographic, clinical, and PN variables.	Image processing and feature extraction performed by expert radiologists. Random forest machine learning algorithm used to select variables and develop CAD model.	Validation cohort: CAD AUC = 0.92. CAD: Sn = 0.95, Sp = 0.88, PPV = 0.86, NPV = 0.96. Three radiologists' combined reading: Sn = 0.70, Sp = 0.69, PPV = 0.64, NPV = 0.75.
Peikert et al, 2018 [55]	Analytical validation study to develop and internally validate a radiomics-based multivariable model (BRODERS model).	Training data: 726 PNs (7–30 mm; 56% malignant) from 726 pts in the NLST.	PNs segmented manually using ANALYZE software (Mayo Clinic Biomedical Imaging Resource) and radiomic features extracted. LASSO multivariable analysis used to develop final model.	Optimism-corrected AUC for final 8-variable BRODERS model = 0.94.
Maldonado et al, 2020 [56]	Analytical validation study to externally validate BRODERS model.	External validation data: 170 PNs (7–30 mm; 54% malignant) from 170 consecutive pts with incidentally detected PNs at Vanderbilt University.	BRODERS model (described above) compared to Brock model.	BRODERS AUC = 0.90; Brock AUC = 0.87.
Balagurunathan et al, 2019 [57]	Analytical validation study using a 2:1 nested case-control study design to develop a novel radiomics model.	Training data: 244 PNs (> 4 mm; 32% malignant) from 244 pts in the NLST. Validation data: 235 PNs (> 4 mm; 37% malignant) from 235 pts in the NLST. Malignant and benign PNs were matched based on demographic and clinical variables.	PNs 3D segmented by radiologists via semi-automated algorithm, 219 quantitative features extracted, and an optimal linear classifier model was used.	In both training (0.85 vs 0.80) and validation (0.88 vs 0.86) datasets, AUC was higher for best texture feature set compared to size and shape feature set. Addition of clinical data did not significantly improve AUC.
Ardila et al, 2019 [58]	Analytical validation study and retrospective reader study to develop and externally validate a novel radiomics-based AI CAD model.	Training data: 29,541 PNs (4% malignant) from NLST. Tuning data: ∼6,343 PNs (5% malignant) from NLST. Validation data: 6,716 PNs (4% malignant) from NLST. Reader study: 6 board-certified radiologists evaluated 507 CTs with PNs (16% malignant; subset of validation data).	CAD approach developed using the TensorFlow platform (Google Inc.) and employed a 3D CNN model that performs end-to-end analysis of whole-CT volumes. Model output = LUMAS, roughly meant to correspond to Lung-RADS 3, 4A, and 4B/4X.	Validation cohort: AI CAD AUC = 0.94. AI CAD outperformed radiologists within each LUMAS bucket in reader study when either 1 CT scan was used per pt or when multiple scans were available per pt.

Table 1, continued

Publication	Study design and objective	Populations or datasets	Model and analytical details	Key results
Venkadesh et al, 2021 [59]	Analytical validation study and retrospective reader study to develop and externally validate a novel radiomics-based AI CAD model.	Training data: 16,077 PNs (> 4 mm; 8% malignant) from NLST. Validation data: 883 PNs in full cohort (7% malignant); 175 non-size-matched PNs in subset A (34% malignant); 177 size-matched PNs in subset B (33% malignant) from the DLCST. Reader study: 11 clinicians (9 radiologists, 2 pulmonologists) evaluated PNs in cancer-enriched cohorts.	2D CNN with ResNet50 backbone and 3D CNN based on Inception-v1 architecture used to develop AI CAD algorithm. Internally validated using 10-fold cross validation. AI CAD model compared to Brock model and clinicians. Model output = risk score from 0 to 1.	Full validation cohort: AI CAD AUC = 0.93; Brock AUC = 0.90. Subset A cohort: AI CAD AUC = 0.96; average clinician AUC = 0.90; Brock AUC = 0.94. Subset B cohort: AI CAD AUC = 0.86; average clinician AUC = 0.82; Brock AUC = 0.75.
Massion et al, 2020 [60]	Analytical validation study to develop and externally validate a novel radiomics-based AI CAD model (Optellum LCP-CNN).	Training data: > 130,000 PNs (∼50% malignant) from NLST. Internal validation data: 15,693 PNs (> 6 mm; 6% malignant) from 6,547 pts in the NLST. External validation data: 116 PNs (5–30 mm; 55% malignant) from 116 pts with incidentally detected PNs at Vanderbilt University; 463 PNs (5–19 mm; 14% malignant) from 427 pts with incidentally detected PNs at Oxford University	2.5D CNN with DenseNet architecture with 5 dense blocks and PyTorch framework for machine learning. Internally validated using 8-fold cross validation. Model output = score between 0% and 100% to represent likelihood of malignancy. Compared to Brock and Mayo Clinic models.	Internal validation cohort: LCP-CNN AUC = 0.92; Brock AUC = 0.86; Mayo Clinic AUC = 0.85. Vanderbilt University external validation cohort: LCP-CNN AUC = 0.84; Mayo Clinic AUC = 0.78. Oxford University external validation cohort: LCP-CNN AUC = 0.92; Mayo Clinic AUC = 0.82.
Baldwin et al, 2020 [61]	Analytical validation study to externally validate the Optellum LCP-CNN model.	External validation data: 1,397 PNs (5–15 mm; 17% malignant) from 1,187 U.K. pts in IDEAL study.	Optellum LCP-CNN model (described above) compared to Brock model.	LCP-CNN AUC = 0.87; Brock AUC = 0.83.
Kim et al, 2022 [62]	Retrospective multi-reader, multi-case study to assess the effect of Optellum AI CAD model on clinicians' performance discriminating between malignant and benign PNs.	Reader study: 12 clinicians (6 radiologists, 6 pulmonologists) evaluated 300 CTs with PNs (5–30 mm; 50% malignant) from 300 pts from 7 sources in the U.S., U.K., and NLST.	Optellum LCP-CNN model (described above). Model output = LCP score 1 to 10, categorizing malignancy risk on a decile scale for a population with 30% cancer prevalence.	Average clinicians' AUC increased from 0.82 to 0.89 with AI CAD. Interobserver agreement (Fleiss Kappa) improved with AI CAD for < 5% risk (0.71 vs 0.50) and > 65% risk (0.71 vs 0.54) categories and PN management decisions (0.52 vs 0.44).
Kim et al, 2023 [63]	Secondary analysis of above retrospective multi-reader, multi-case study to assess the effect of Optellum AI CAD model on clinicians' management of PNs.	Reader study: described above.	LCP score (described above). Appropriate PN management defined as surgery, biopsy, or immediate imaging for malignant PNs and imaging follow-up for benign PNs.	Average clinicians' risk estimate without vs with AI CAD: 60% vs 69% (malignant PNs); 23% vs 21% (benign PNs). Average clinicians' appropriate PN management without vs with AI CAD: 80% vs 84% (overall); 72% vs 81% (malignant PNs); 87% vs 89% (benign PNs).

Abbreviations: CAD = computer-aided diagnosis; PN = pulmonary nodule; pts = patients; LIDC = Lung Image Database Consortium; LDA = linear discriminant analysis; NLST = National Lung Screening Trial; Sn = sensitivity; Sp = specificity; PPV = positive predictive value; NPV = negative predictive value; BRODERS = Benign Versus Aggressive Nodule Evaluation Using Radiomics Stratification; LASSO = least absolute shrinkage and selection operator; AI = artificial intelligence; CNN = convolutional neural network; CT = computed tomography; LUMAS = lung malignancy score; Lung-RADS = Lung Imaging reporting and Data System; DLCST = Danish Lung Cancer Screening Trial; LCP-CNN = Lung Cancer Prediction Convolutional Neural Network; U.K. = United Kingdom; IDEAL = Artificial Intelligence and Big Data for Early Lung Cancer Diagnosis; U.S. = United States.

specificity of 0.88, positive predictive value (PPV) of 0.86, and a negative predictive value (NPV) of 0.96, which outperformed three radiologists' collective evaluations (sensitivity: 0.70, specificity: 0.69, PPV: 0.64, NPV: 0.75). In 2018, Peikert and colleagues also used NLST data to develop a distinct radiomics-based model via manual software segmentation, incorporation of both PN and adjacent lung tissue characteristics, and the least absolute shrinkage selection operator (LASSO) method for multivariable model development, and reported an associated AUC of 0.94 on internal validation [55]. Subsequent external validation of this model using data from the Vanderbilt University Lung Nodule Registry yielded an AUC of 0.90 [56]. In 2019, Balagurunathan and colleagues published the results of their radiomics-based models also trained on NLST data reporting an AUC as high as 0.85 and noting the superior contribution of texture metrics in comparison to traditional size metrics [57]. The authors also found that discrimination was not augmented when clinical factors were incorporated into their radiomics-based models.

An alternative method of harnessing and analyzing radiomics-based quantitative imaging data from CT scans to develop a predictive model requires the use of AI [48,49]. Advancements in AI have ushered in the emergence of deep learning algorithms that do not rely on explicit feature parameter inputs but instead are trained via direct interaction with the data, theoretically enhancing problem-solving abilities. Convolutional neural networks (CNNs) are currently the most commonly used deep learning architecture in medical imaging. Generally speaking, these AI deep learning algorithms simultaneously evaluate imaging data, extract and aggregate features, and integrate this information to achieve high-level reasoning and ultimately make a prediction regarding PN malignancy risk. Radiomics-based tools that use AI technology fundamentally differ from those that do not, as these algorithms "learn" independently, can potentially identify previously unknown imaging features, and are capable of being iteratively updated by the introduction of new training data. A small but growing number of radiomics-based AI tools have been developed to date. In 2019, Ardila and colleagues described the development of a CNN model designed by Google that was trained on and validated in NLST imaging data. Notably, this model used full-volume imaging data (i.e,. the entire axial series of images) to classify malignancy risk. They reported an AUC of 0.94, which outperformed six radiologists [58]. The authors proposed a four-tier lung malignancy scoring (LUMAS) system, loosely meant to correspond with estimated malignancy probabilities associated with American College of Radiology Lung-RADS categories, but emphasized that optimization of this scoring system for use in clinical practice had yet to be performed. Separately, in 2021 Venkadesh and colleagues published the results of their CNN-based algorithm that was trained on NLST data and externally validated using data from the Danish Lung Cancer Screening Trial. Their deep learning algorithm outperformed the Brock (PanCan) traditional clinical risk prediction model (AUC: 0.93 vs 0.90; $P < 0.05$) and performed similarly to thoracic radiologists (AUC: 0.96 vs 0.90; $P = 0.11$) [59]. The authors initially made their algorithm freely accessible to the public for a time and concluded that their AI-based algorithm could serve as an adjunct for radiologists evaluating screening CT scans in the future.

To date, the only radiomics-based AI algorithm to gain both U.S. Food and Drug Administration 510(k) clearance (2021) and European Union CE marking (2022) is the Lung Cancer Prediction Convolutional Neural Network (LCP-CNN) developed by Optellum. This AI CAD tool was trained on and internally validated in NLST data of screen-detected PNs (AUC: 0.92) and was externally validated using imaging data of incidentally-detected PNs from Vanderbilt University Medical Center (AUC: 0.84), Oxford University Hospital National Health Service (NHS) Foundation Trust (AUC: 0.92), Leeds Teaching Hospital NHS Trust (AUC: 0.88), and Nottingham University Hospitals NHS Trust (AUC: 0.89) [60,61]. Additionally, the LCP-CNN had superior discrimination compared to both the Mayo Clinic and the Brock (PanCan) clinical models. A commercially available version of the LCP-CNN generates a radiomics biomarker Lung Cancer Prediction (LCP) score that represents an estimate of predicted risk of malignancy on a decile scale. In 2022, a retrospective multi-reader, multi-case study was performed to evaluate the effect of the LCP-CNN on clinicians' malignancy risk assessments [62]. Twelve clinicians (six pulmonologists and six radiologists) each evaluated 300 chest CT cases of PNs and were asked to provide an estimate of PN malignancy risk (0%–100%) and a management recommendation for each case before and after using the AI tool. When using the tool, clinicians' average discrimination improved by 7 percentage points (AUC: 0.89 vs 0.82; $P < 0.001$) and sensitivity and specificity at both the 5% and 65% malignancy risk thresholds increased as well. Interobserver agreement for both clinically relevant malignancy risk categories

($< 5\%$, 5%–30%, 31%–65%, $> 65\%$) and management recommendations (no action, CT surveillance, diagnostic procedure) also increased with use of the AI tool. Moreover, the average proportion of appropriately managed PN cases (defined as immediate imaging or biopsy for malignant PNs and no action or imaging surveillance for benign PNs) increased from 80% to 84% with use of the LCP-CNN in this retrospective study [63].

3. Barriers to implementation

Despite the plethora of novel radiomics and AI-based CAD tools that have been developed and the well-known need for improved PN risk stratification, widespread adoption of this technology has not yet occurred despite being commercially available. The reason why is likely multifaceted. First, while all of the aforementioned studies reported metrics for model performance (i.e., AUC, sensitivity, and specificity), prospective clinical utility studies using real-world data have not yet been performed. It is critical to note that models associated with high levels of discrimination (i.e., AUC) do not necessarily equate to high-performing models in clinical settings that differ from patient populations in which models were originally trained and validated [64]. Specifically, differences in demographic characteristics and cancer prevalence could limit generalizability of model performance in distinct populations. In fact, the more relevant metric for model performance and applicability to specific patient care scenarios is model calibration [65]. Currently, there does not exist a standardized approach to systematically evaluate AI in healthcare or how best to evaluate the clinical utility of new technologies. However, several approaches to rigorously evaluating novel AI technologies have been proposed. For example, Park and colleagues have proposed an approach akin to the classic framework for new drug development, advancing scientific inquiry from phase 1 safety-focused studies to eventual phase 4 clinical effectiveness studies [66]. Khera and colleagues have suggested a holistic approach to AI evaluation and implementation with an emphasis on health quality, equity, generalizability, and medical education in addition to evaluating patient-centered outcomes [67]. Of course, the optimal method for evaluating any novel intervention is to perform a prospective randomized controlled trial assessing patient-centered outcomes. To date, no such studies have been published. Second, much has been made of the unique challenges AI technology poses in the medical setting. As AI tools use an automated approach to independent learning, concerns have been raised regarding the "black-box" nature of which factors drive AI decision-making and risk estimation [68]. This opaqueness in what is "under the hood" of AI algorithms have resulted in mistrust among clinicians [69]. In fact, a recent survey of clinicians highlighted limited acceptance and trust of AI technology as a significant perceived barrier to implementation [70]. This survey also revealed clinicians' concerns about safety, inconsistent technical performance, absence of standardized guidelines, lack of technical knowledge, and loss of autonomy. Radiologists have additionally raised concerns regarding medical-legal liability, responsibility for the results of AI-generated recommendations, and the nature of AI integration into routine clinical workflow [71]. Third, as radiomics-based tools require high resolution CT images to be available and large imaging data files to be uploaded into CAD software platforms, practical barriers to clinical implementation include lack of standardization of CT image acquisition across different healthcare institutions and disruption of clinical workflow in already busy pulmonary nodule clinics. Finally, the medical community's overall wariness of AI technology is understandable given previous examples of unintended consequences of CAD on medical decision-making [72,73]. For example, a 2003 study assessing the effect of CAD on electrocardiogram (ECG) interpretation by inexperienced resident physicians demonstrated that when incorrect CAD interpretations were provided, residents were more likely to misinterpret an ECG compared to when CAD was not used [74]. In another study, use of CAD was associated with a reduction in breast cancer discrimination on mammography among high-performing expert clinicians [75]. Subsequent studies reported either no significant impact of CAD on radiologists' decision-making [76] or a decrease in clinician discrimination when using CAD [77]. These examples underscore the importance and need to perform high quality studies assessing the effect of CAD tools on both clinical decision-making and patient outcomes.

4. Future directions

Before widespread implementation of radiomics-based AI tools for PN risk stratification can be recommended, well-executed studies must be performed to assess the effect of such tools on medical decision-making and patient-centered outcomes and to determine how best to implement these devices into routine clinical practice. Importantly, AI algorithms have been developed and trained to discriminate between malignant and benign PNs, but they are not capable of understanding the nuances of patient preferences and clinician assessments of the associated risks of various management approaches [68,69]. For a given indeterminate PN, clinicians have inconsistent approaches to PN risk assessment and variable malignancy probability thresholds above which they would recommend pursuing a lung biopsy [29,78,79,80,81]. For example, a more conservative clinician might not recommend a biopsy unless a PN diameter is greater than 10 mm or unless the estimated malignancy risk is greater than 20% or 30%, whereas a more aggressive approach might see a clinician recommend a biopsy for any PN larger than 8 mm or with a risk greater than 10%. Apart from clinicians' variable perspectives on PN malignancy risk and management, individual patients can have widely disparate opinions on acceptable risk and anxiety related to the lack of certainty associated with a PN detected on a CT scan [82,83,84]. For example, a patient who values not missing a cancer diagnosis and places high importance on timeliness of care might choose to pursue a biopsy upfront for a given indeterminate PN even at the lower end of malignancy risk. On the other hand, a patient with multiple comorbidities who might be more anxious of the potential risks and complications of a lung biopsy procedure might choose to avoid a biopsy initially, opting for surveillance with serial CT scans instead. Thus, radiomics-based AI tools are not designed to replace clinicians' decision-making but, at best, could assist clinicians and patients in jointly making the challenging decision of whether or not to biopsy a given PN [85]. As such, several decision analytic modeling approaches to estimating the clinical utility of diagnostic tests that take into account various threshold probabilities for biopsy have been developed. The most widely used and oldest is decision curve analysis, developed by Vickers and colleagues in 2006 [86,87,88,89]. This analytic technique plots net clinical benefit (a weighted difference between true positives and false positives for malignancy) on the Y-axis against threshold probability (the malignancy probability above which biopsy would be recommended) on the X-axis and has been used in multiple areas of research [90,91,92]. Notable examples of alternative approaches include the relative clinical utility curve developed by Baker and colleagues [93, 94,95,96] and the interventional probability curve from Kammer and colleagues [97]. A necessary first step to understanding the potential effect of novel AI tools on PN management decisions will be the rigorous application of such clinical utility models using real-world patient data.

Promisingly, a growing number of studies have begun to estimate the clinical utility of radiomics-based tools in a retrospective fashion. For example, a recent publication from Paez and colleagues demonstrated the potential clinical utility of the Optellum LCP-CNN for longitudinal assessment of PNs, as malignancy risk estimates for malignant PNs increased over time while those for benign PNs remained relatively stable [98]. Separately, in 2021 Kammer and colleagues described the development of a novel combination biomarker incorporating clinical variables in addition to blood and radiomics-based inputs and performed a clinical utility analysis to estimate what the effect of using the biomarker would have been on clinical decision-making [99]. They found that use of this novel biomarker would theoretically have both reduced the proportion of individuals with benign PNs undergoing invasive procedures and the time to diagnosis of cancer among those with malignant PNs.

As previously mentioned, the gold standard method of evaluating any novel intervention is to perform a prospective randomized controlled trial that directly assesses the impact of an intervention on patient-centered clinical outcomes. Multiple experts have urged the performance of such trials when evaluating any novel AI-based technology [69,72,100,101]. To date, no clinical trials have been conducted evaluating the clinical effectiveness of a radiomics-based AI tool on PN risk stratification. However, a recent search of ClinicalTrials.gov reveals one such trial that is actively recruiting patients (NCT05968898). This pragmatic randomized controlled trial will compare usual care with an approach to PN risk stratification that incorporates use of the Optellum LCP-CNN tool. The primary outcome will be the composite proportion of malignant PNs managed with biopsy or empiric treatment and benign PNs managed with imaging surveillance, and secondary outcomes include timeliness of care, adverse events, diagnostic yield of biopsy procedures, and healthcare costs. Thus, much needed future efforts to carefully investigate AI technology are currently in the pipeline.

5. Conclusions

In conclusion, recent advances in radiomics-based AI technology have yielded promising preliminary data suggesting that AI may serve a complementary role to routine clinical decision-making for PN management in the future. However, widespread adoption of such novel tools has not yet been observed despite commercial availability, and use of such technology is not currently recommended by any clinical guidelines due to a dearth of adequate clinical utility and prospective randomized controlled trial data. Future rigorously conducted clinical research studies are required to fully evaluate the clinical effectiveness of radiomics-based AI tools for PN risk stratification and to clearly define what role, if any, these tools should play within routine clinical practice.

Author contributions

R.Y.K. performed the literature review and wrote the manuscript.

Funding

R.Y.K. is supported by a National Cancer Institute Career Development Award (K08CA279881).

Conflict of interest

No relevant financial conflicts of interest to disclose.

References

[1] R. Smith-Bindman, D.L. Miglioretti, E. Johnson, C. Lee, H.S. Feigelson, M. Flynn, R.T. Greenlee, R.L. Kruger, M.C. Hornbrook, D. Roblin, L.I. Solberg, N. Vanneman, S. Weinmann and A.E. Williams, Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996–2010, *JAMA* **307** (2012), 2400–2409.

[2] M.K. Gould, T. Tang, I.L. Liu, J. Lee, C. Zheng, K.N. Danforth, A.E. Kosco, J.L. Di Fiore and D.E. Suh, Recent trends in the identification of incidental pulmonary nodules, *Am J Respir Crit Care Med* **192** (2015), 1208–1214.

[3] V.A. Moyer and U.S.P.S.T. Force, Screening for lung cancer: U.S. Preventive Services Task Force recommendation statement, *Ann Intern Med* **160** (2014), 330–338.

[4] R. Meza, J. Jeon, I. Toumazis, K. Ten Haaf, P. Cao, M. Bastani, S.S. Han, E.F. Blom, D.E. Jonas, E.J. Feuer, S.K. Plevritis, H.J. de Koning and C.Y. Kong, Evaluation of the benefits and harms of lung cancer screening with low-dose computed tomography: Modeling study for the US Preventive Services Task Force, *JAMA* **325** (2021), 988–997.

[5] R.L. Siegel, K.D. Miller, H.E. Fuchs and A. Jemal, Cancer statistics, 2022, *CA Cancer J Clin* **72** (2022), 7–33.

[6] J. Huo, Y. Xu, T. Sheu, R.J. Volk and Y.T. Shih, Complication rates and downstream medical costs associated with invasive diagnostic procedures for lung abnormalities in the community setting, *JAMA Intern Med* **179** (2019), 324–332.

[7] H. Zhao, Y. Xu, J. Huo, A.C. Burks, D.E. Ost and Y.T. Shih, Updated analysis of complication rates associated with invasive diagnostic procedures after lung cancer screening, *JAMA Netw Open* **3** (2020), e2029874.

[8] S.P.E. Nishi, J. Zhou, I. Okereke, Y.F. Kuo and J. Goodwin, Use of imaging and diagnostic procedures after Low-Dose CT screening for lung cancer, *Chest* **157** (2020), 427–434.

[9] F. Farjah, S.E. Monsell, M.K. Gould, R. Smith-Bindman, M.P. Banegas, P.J. Heagerty, E.M. Keast, A. Ramaprasan, K. Schoen, E.G. Brewer, R.T. Greenlee and D.S.M. Buist, Association of the intensity of diagnostic evaluation with outcomes in incidentally detected lung nodules, *JAMA Intern Med* **181** (2021), 480–489.

[10] K.A. Rendle, C.A. Saia, A. Vachani, A.N. Burnett-Hartman, V.P. Doria-Rose, S. Beucker, C. Neslund-Dudas, C. Oshiro, R.Y. Kim, J. Elston-Lafata, S.A. Honda, D. Ritzwoller, J.V. Wainwright, N. Mitra and R.T. Greenlee, Rates of downstream procedures and complications associated with lung cancer screening in routine clinical practice: A retrospective cohort study, *Ann Intern Med* **177** (2024), 18–28.

[11] D.E. Ost and M.K. Gould, Decision making in patients with pulmonary nodules, *Am J Respir Crit Care Med* **185** (2012), 363–372.

[12] M.K. Gould, J. Donington, W.R. Lynch, P.J. Mazzone, D.E. Midthun, D.P. Naidich and R.S. Wiener, Evaluation of individuals with pulmonary nodules: when is it lung cancer? Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines, *Chest* **143** (2013), e93S–e120S.

[13] H. MacMahon, D.P. Naidich, J.M. Goo, K.S. Lee, A.N.C. Leung, J.R. Mayo, A.C. Mehta, Y. Ohno, C.A. Powell, M. Prokop, G.D. Rubin, C.M. Schaefer-Prokop, W.D. Travis, P.E. Van Schil and A.A. Bankier, Guidelines for management of incidental pulmonary nodules detected on CT images: From the Fleischner Society 2017, *Radiology* **284** (2017), 228–243.

[14] American College of Radiology, Lung CT Screening Reporting and Data System (Lung-RADS), in.

[15] N.T. Tanner, A. Porter, M.K. Gould, X.J. Li, A. Vachani and G.A. Silvestri, Physician assessment of pretest probability of malignancy and adherence with guidelines for pulmonary nodule evaluation, *Chest* **152** (2017), 263–270.

[16] N.T. Tanner, J. Aggarwal, M.K. Gould, P. Kearney, G. Diette, A. Vachani, K.C. Fang and G.A. Silvestri, Management of pulmonary nodules by community pulmonologists: A multicenter observational study, *Chest* **148** (2015), 1405–1414.

[17] T. Lokhandwala, M.A. Bittoni, R.A. Dann, A.O. D'Souza, M. Johnson, R.J. Nagy, R.B. Lanman, R.E. Merritt and D.P. Carbone, Costs of diagnostic assessment for lung cancer: A medicare claims analysis, *Clin Lung Cancer* **18** (2017), e27–e34.

[18] R.S. Wiener, M.K. Gould, C.G. Slatore, B.G. Fincke, L.M. Schwartz and S. Woloshin, Resource use and guideline concordance in evaluation of pulmonary nodules for cancer: too much and too little care, *JAMA Intern Med* **174** (2014), 871–880.

[19] G.A. Silvestri, A. Vachani, D. Whitney, M. Elashoff, K. Porta Smith, J.S. Ferguson, E. Parsons, N. Mitra, J. Brody, M.E. Lenburg, A. Spira and A.S. Team, A bronchial genomic classifier for the diagnostic evaluation of lung cancer, *N Engl J Med* **373** (2015), 243–251.

[20] T. National Lung Screening Trial Research, D.R. Aberle, A.M. Adams, C.D. Berg, W.C. Black, J.D. Clapp, R.M. Fagerstrom, I.F. Gareen, C. Gatsonis, P.M. Marcus and J.D. Sicks,

Reduced lung-cancer mortality with low-dose computed tomographic screening, *N Engl J Med* **365** (2011), 395–409.

[21] S.J. Swensen, M.D. Silverstein, D.M. Ilstrup, C.D. Schleck and E.S. Edell, The probability of malignancy in solitary pulmonary nodules, *Archives of Internal Medicine* **157** (1997).

[22] G.J. Herder, H. van Tinteren, R.P. Golding, P.J. Kostense, E.F. Comans, E.F. Smit and O.S. Hoekstra, Clinical prediction model to characterize pulmonary nodules: validation and added value of 18F-fluorodeoxyglucose positron emission tomography, *Chest* **128** (2005), 2490–2496.

[23] M. Reid, H.K. Choi, X. Han, X. Wang, S. Mukhopadhyay, L. Kou, U. Ahmad, X. Wang and P.J. Mazzone, Development of a risk prediction model to estimate the probability of malignancy in pulmonary nodules being considered for biopsy, *Chest* **156** (2019), 367–375.

[24] A. McWilliams, M.C. Tammemagi, J.R. Mayo, H. Roberts, G. Liu, K. Soghrati, K. Yasufuku, S. Martel, F. Laberge, M. Gingras, S. Atkar-Khattra, C.D. Berg, K. Evans, R. Finley, J. Yee, J. English, P. Nasute, J. Goffin, S. Puksa, L. Stewart, S. Tsai, M.R. Johnston, D. Manos, G. Nicholas, G.D. Goss, J.M. Seely, K. Amjadi, A. Tremblay, P. Burrowes, P. MacEachern, R. Bhatia, M.S. Tsao and S. Lam, Probability of cancer in pulmonary nodules detected on first screening CT, *N Engl J Med* **369** (2013), 910–919.

[25] H. MacMahon, F. Li, Y. Jiang and S.G. Armato, 3rd, Accuracy of the vancouver lung cancer risk prediction model compared with that of radiologists, *Chest* **156** (2019), 112–119.

[26] A.A. Balekian, G.A. Silvestri, S.M. Simkovich, P.J. Mestaz, G.D. Sanders, J. Daniel, J. Porcel and M.K. Gould, Accuracy of clinicians and models for estimating the probability that a pulmonary nodule is malignant, *Ann Am Thorac Soc* **10** (2013), 629–635.

[27] N.T. Tanner, P.B. Brasher, J. Jett and G.A. Silvestri, Effect of a rule-in biomarker test on pulmonary nodule management: A survey of pulmonologists and thoracic surgeons, *Clin Lung Cancer* **21** (2020), e89–e98.

[28] A.W. Maiga, S.A. Deppen, P.P. Massion, C. Callaway-Lane, R. Pinkerman, R.S. Dittus, E.S. Lambright, J.C. Nesbitt and E.L. Grogan, Communication about the probability of cancer in indeterminate pulmonary nodules, *JAMA Surg* **153** (2018), 353–357.

[29] J.M. Iaccarino, J. Simmons, M.K. Gould, C.G. Slatore, S. Woloshin, L.M. Schwartz and R.S. Wiener, Clinical equipoise and shared decision-making in pulmonary nodule management. A survey of American Thoracic Society Clinicians, *Ann Am Thorac Soc* **4** (2017), 968–975.

[30] C.G. Slatore, N. Horeweg, J.R. Jett, D.E. Midthun, C.A. Powell, R.S. Wiener, J.P. Wisnivesky, M.K. Gould and A.T.S.A.H.C.o.S.a.R.F.f.P.N. Evaluation, An official American Thoracic Society research statement: A research framework for pulmonary nodule evaluation and management, *Am J Respir Crit Care Med* **192** (2015), 500–514.

[31] R. Paez, M.N. Kammer, N.T. Tanner, S. Shojaee, B.E. Heideman, T. Peikert, M.L. Balbach, W.T. Iams, B. Ning, M.E. Lenburg, C. Mallow, L. Yarmus, K.M. Fong, S. Deppen, E.L. Grogan and F. Maldonado, Update on biomarkers for the stratification of indeterminate pulmonary nodules, *Chest* (2023).

[32] M.N. Kammer and P.P. Massion, Noninvasive biomarkers for lung cancer diagnosis, where do we stand? *J Thorac Dis* **12** (2020), 3317–3330.

[33] H. Mamdani, S. Ahmed, S. Armstrong, T. Mok and S.I. Jalal, Blood-based tumor biomarkers in lung cancer for detection and treatment, *Transl Lung Cancer Res* **6** (2017), 648–660.

[34] R. Tao, W. Cao, F. Zhu, J. Nie, H. Wang, L. Wang, P. Liu, H. Chen, B. Hong and D. Zhao, Liquid biopsies to distinguish malignant from benign pulmonary nodules, *Thorac Cancer* **12** (2021), 1647–1655.

[35] C. Liu, X. Xiang, S. Han, H.Y. Lim, L. Li, X. Zhang, Z. Ma, L. Yang, S. Guo, R. Soo, B. Ren, L. Wang and B.C. Goh, Blood-based liquid biopsy: Insights into early detection and clinical management of lung cancer, *Cancer Lett* **524** (2022), 91–102.

[36] G.A. Silvestri, N.T. Tanner, P. Kearney, A. Vachani, P.P. Massion, A. Porter, S.C. Springmeyer, K.C. Fang, D. Midthun, P.J. Mazzone and P.T. Team, Assessment of plasma proteomics biomarker's ability to distinguish benign from malignant lung nodules: Results of the PANOPTIC (Pulmonary Nodule Plasma Proteomic Classifier) trial, *Chest* **154** (2018), 491–500.

[37] A. Vachani, D.H. Whitney, E.C. Parsons, M. Lenburg, J.S. Ferguson, G.A. Silvestri and A. Spira, Clinical utility of a bronchial genomic classifier in patients with suspected lung cancer, *Chest* **150** (2016), 210–218.

[38] A.S. Team, Shared gene expression alterations in nasal and bronchial epithelium for lung cancer detection, *J Natl Cancer Inst* **109** (2017).

[39] R.J. Keogh and J.C. Riches, The use of breath analysis in the management of lung cancer: Is it ready for primetime? *Curr Oncol* **29** (2022), 7355–7378.

[40] I. Horvath, Z. Lazar, N. Gyulai, M. Kollai and G. Losonczy, Exhaled biomarkers in lung cancer, *Eur Respir J* **34** (2009), 261–275.

[41] P. Wang, Q. Huang, S. Meng, T. Mu, Z. Liu, M. He, Q. Li, S. Zhao, S. Wang and M. Qiu, Identification of lung cancer breath biomarkers based on perioperative breathomics testing: A prospective observational study, *EClinicalMedicine* **47** (2022), 101384.

[42] Y.J. Wu, F.Z. Wu, S.C. Yang, E.K. Tang and C.H. Liang, Radiomics in early lung cancer diagnosis: From diagnosis to clinical decision support and education, *Diagnostics (Basel)* **12** (2022).

[43] A. Khawaja, B.J. Bartholmai, S. Rajagopalan, R.A. Karwoski, C. Varghese, F. Maldonado and T. Peikert, Do we need to see to believe?-radiomics for lung nodule classification and lung cancer risk stratification, *J Thorac Dis* **12** (2020), 3303–3316.

[44] H. Fujita, AI-based computer-aided diagnosis (AI-CAD): the latest review to read first, *Radiol Phys Technol* **13** (2020), 6–19.

[45] R.J. Gillies, P.E. Kinahan and H. Hricak, Radiomics: Images are more than pictures, *They Are Data, Radiology* **278** (2016), 563–577.

[46] R. Wilson and A. Devaraj, Radiomics of pulmonary nodules and lung cancer, *Transl Lung Cancer Res* **6** (2017), 86–91.

[47] Y. Yang, X. Feng, W. Chi, Z. Li, W. Duan, H. Liu, W. Liang, W. Wang, P. Chen, J. He and B. Liu, Deep learning aided decision support for pulmonary nodules diagnosing: A review, *J Thorac Dis* **10** (2018), S867–S875.

[48] A. Hosny, C. Parmar, J. Quackenbush, L.H. Schwartz and H. Aerts, Artificial intelligence in radiology, *Nat Rev Cancer* **18** (2018), 500–510.

[49] S. Ather, T. Kadir and F. Gleeson, Artificial intelligence and radiomics in pulmonary nodule management: Current status and future applications, *Clin Radiol* **75** (2020), 13–19.

[50] W. Lv, Y. Wang, C. Zhou, M. Yuan, M. Pang, X. Fang, Q. Zhang, C. Huang, X. Li, Z. Zhou, Y. Yu, Y. Wang, M. Lu, Q. Xu, X. Li, H. Lin, X. Lu, Q. Xu, J. Sun, Y. Tang, F. Yan, B. Zhang, Z. Cheng, L. Zhang and G. Lu, Development and

validation of a clinically applicable deep learning strategy (HONORS) for pulmonary nodule classification at CT: A retrospective multicentre study, *Lung Cancer* **155** (2021), 78–86.

[51] T.W. Way, L.M. Hadjiiski, B. Sahiner, H.P. Chan, P.N. Cascade, E.A. Kazerooni, N. Bogot and C. Zhou, Computer-aided diagnosis of pulmonary nodules on CT scans: segmentation and classification using 3D active contours, *Med Phys* **33** (2006), 2323–2337.

[52] T.W. Way, B. Sahiner, H.P. Chan, L. Hadjiiski, P.N. Cascade, A. Chughtai, N. Bogot and E. Kazerooni, Computer-aided diagnosis of pulmonary nodules on CT scans: improvement of classification performance with nodule surface features, *Med Phys* **36** (2009), 3086–3098.

[53] T. Way, H.P. Chan, L. Hadjiiski, B. Sahiner, A. Chughtai, T.K. Song, C. Poopat, J. Stojanovska, L. Frank, A. Attili, N. Bogot, P.N. Cascade and E.A. Kazerooni, Computer-aided diagnosis of lung nodules on CT scans: ROC study of its effect on radiologists' performance, *Acad Radiol* **17** (2010), 323–332.

[54] P. Huang, S. Park, R. Yan, J. Lee, L.C. Chu, C.T. Lin, A. Hussien, J. Rathmell, B. Thomas, C. Chen, R. Hales, D.S. Ettinger, M. Brock, P. Hu, E.K. Fishman, E. Gabrielson and S. Lam, Added value of computer-aided CT image features for early lung cancer diagnosis with small pulmonary nodules: A matched case-control study, *Radiology* **286** (2018), 286–295.

[55] T. Peikert, F. Duan, S. Rajagopalan, R.A. Karwoski, R. Clay, R.A. Robb, Z. Qin, J. Sicks, B.J. Bartholmai and F. Maldonado, Novel high-resolution computed tomography-based radiomic classifier for screen-identified pulmonary nodules in the National Lung Screening Trial, *PLoS ONE* **13** (2018), e0196910.

[56] F. Maldonado, C. Varghese, S. Rajagopalan, F. Duan, A. Balar, D.A. Lakhani, S.B. Antic, P. Massion, T.F. Johnson, R.A. Karwoski, R.A. Robb, B.J. Bartholmai and T. Peikert, Validation of the BRODERS classifier (Benign versus aggressive nODule Evaluation using Radiomic Stratification), a novel high-resolution computed tomography-based radiomic classifier for indeterminate pulmonary nodules, *Eur Respir J* (2020).

[57] Y. Balagurunathan, M.B. Schabath, H. Wang, Y. Liu and R.J. Gillies, Quantitative imaging features improve discrimination of malignancy in pulmonary nodules, *Sci Rep* **9** (2019), 8528.

[58] D. Ardila, A.P. Kiraly, S. Bharadwaj, B. Choi, J.J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, D.P. Naidich and S. Shetty, End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography, *Nat Med* **25** (2019), 954–961.

[59] K.V. Venkadesh, A.A.A. Setio, A. Schreuder, E.T. Scholten, K. Chung, W.W. MM, Z. Saghir, B. van Ginneken, M. Prokop and C. Jacobs, Deep learning for malignancy risk estimation of pulmonary nodules detected at Low-Dose Screening CT, *Radiology* (2021), 204433.

[60] P.P. Massion, S. Antic, S. Ather, C. Arteta, J. Brabec, H. Chen, J. Declerck, D. Dufek, W. Hickes, T. Kadir, J. Kunst, B.A. Landman, R.F. Munden, P. Novotny, H. Peschl, L.C. Pickup, C. Santos, G.T. Smith, A. Talwar and F. Gleeson, Assessing the accuracy of a deep learning method to risk stratify indeterminate pulmonary nodules, *Am J Respir Crit Care Med* **202** (2020), 241–249.

[61] D.R. Baldwin, J. Gustafson, L. Pickup, C. Arteta, P. Novotny, J. Declerck, T. Kadir, C. Figueiras, A. Sterba, A. Exell, V. Potesil, P. Holland, H. Spence, A. Clubley, E. O'Dowd, M. Clark, V. Ashford-Turner, M.E. Callister and F.V. Gleeson, External validation of a convolutional neural network artificial intelligence tool to predict malignancy in pulmonary nodules, *Thorax* **75** (2020), 306–312.

[62] R.Y. Kim, J.L. Oke, L.C. Pickup, R.F. Munden, T.L. Dotson, C.R. Bellinger, A. Cohen, M.J. Simoff, P.P. Massion, C. Filippini, F.V. Gleeson, and A. Vachani, Artificial intelligence tool for assessment of indeterminate pulmonary nodules detected with CT, *Radiology* **304** (2022), 683–691.

[63] R.Y. Kim, J.L. Oke, T.L. Dotson, C.R. Bellinger and A. Vachani, Effect of an artificial intelligence tool on management decisions for indeterminate pulmonary nodules, *Respirology* **28** (2023), 582–584.

[64] A.A.H. de Hond, E.W. Steyerberg and B. van Calster, Interpreting area under the receiver operating characteristic curve, *Lancet Digit Health* **4** (2022), e853–e855.

[65] B. Van Calster, E.W. Steyerberg, L. Wynants and M. van Smeden, There is no such thing as a validated prediction model, *BMC Med* **21** (2023), 70.

[66] Y. Park, G.P. Jackson, M.A. Foreman, D. Gruen, J. Hu and A.K. Das, Evaluating artificial intelligence in medicine: phases of clinical research, *JAMIA Open* **3** (2020), 326–331.

[67] R. Khera, A.J. Butte, M. Berkwits, Y. Hswen, A. Flanagin, H. Park, G. Curfman and K. Bibbins-Domingo, AI in medicine-JAMA's focus on clinical outcomes, patient-centered care, quality, and equity, *JAMA* (2023).

[68] R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards and K. Tsaneva-Atanasova, Artificial intelligence, bias and clinical safety, *BMJ Qual Saf* **28** (2019), 231–237.

[69] K.H. Yu and I.S. Kohane, Framing the challenges of artificial intelligence in medicine, *BMJ Qual Saf* **28** (2019), 238–241.

[70] L. Strohm, C. Hehakaya, E.R. Ranschaert, W.P.C. Boon and E.H.M. Moors, Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors, *Eur Radiol* **30** (2020), 5525–5532.

[71] E. Neri, F. Coppola, V. Miele, C. Bibbolino and R. Grassi, Artificial intelligence: Who is responsible for the diagnosis? *Radiol Med* **125** (2020), 517–521.

[72] F. Cabitza, R. Rasoini and G.F. Gensini, Unintended consequences of machine learning in medicine, *JAMA* **318** (2017), 517–518.

[73] A. Kohli and S. Jha, Why CAD failed in mammography, *J Am Coll Radiol* **15** (2018), 535–537.

[74] T.L. Tsai, D.B. Fridsma and G. Gatti, Computer decision support as a source of interpretation error: the case of electrocardiograms, *J Am Med Inform Assoc* **10** (2003), 478–483.

[75] A.A. Povyakalo, E. Alberdi, L. Strigini and P. Ayton, How to discriminate between computer-aided and computer-hindered decisions: a case study in mammography, *Med Decis Making* **33** (2013), 98–107.

[76] E.B. Cole, Z. Zhang, H.S. Marques, R. Edward Hendrick, M.J. Yaffe and E.D. Pisano, Impact of computer-aided detection systems on radiologist accuracy with digital mammography, *AJR Am J Roentgenol* **203** (2014), 909–916.

[77] C.D. Lehman, R.D. Wellman, D.S. Buist, K. Kerlikowske, A.N. Tosteson, D.L. Miglioretti and C. Breast Cancer Surveillance, Diagnostic accuracy of digital screening mammography with and without computer-aided detection, *JAMA Intern Med* **175** (2015), 1828–1837.

[78] A. Nair, E.C. Bartlett, S.L.F. Walsh, A.U. Wells, N. Navani, G. Hardavella, S. Bhalla, L. Calandriello, A. Devaraj, J.M. Goo, J.S. Klein, H. MacMahon, C.M. Schaefer-Prokop, J.B. Seo, N. Sverzellati, S.R. Desai, G. Lung Nodule Evaluation and G. Lung Nodule Evaluation, Variable radiological lung nodule evaluation leads to divergent management recommen-

dations, *Eur Respir J* **52** (2018).

[79] F.C. Verdial, D.K. Madtes, G.S. Cheng, S. Pipavath, R. Kim, J.J. Hubbard, M. Zadworny, D.E. Wood and F. Farjah, Multidisciplinary team-based management of incidentally detected lung nodules, *Chest* **157** (2020), 985–993.

[80] S.J. van Riel, C.I. Sanchez, A.A. Bankier, D.P. Naidich, J. Verschakelen, E.T. Scholten, P.A. de Jong, C. Jacobs, E. van Rikxoort, L. Peters-Bax, M. Snoeren, M. Prokop, B. van Ginneken and C. Schaefer-Prokop, Observer variability for classification of pulmonary nodules on Low-Dose CT images and its effect on nodule management, *Radiology* **277** (2015), 863–871.

[81] S.J. van Riel, C. Jacobs, E.T. Scholten, R. Wittenberg, M.M. Winkler Wille, B. de Hoop, R. Sprengers, O.M. Mets, B. Geurts, M. Prokop, C. Schaefer-Prokop and B. van Ginneken, Observer variability for Lung-RADS categorisation of lung cancer screening CTs: impact on patient management, *Eur Radiol* **29** (2019), 924–931.

[82] R.S. Wiener, M.K. Gould, S. Woloshin, L.M. Schwartz and J.A. Clark, What do you mean, a spot? A qualitative analysis of patients' reactions to discussions with their physicians about pulmonary nodules, *Chest* **143** (2013), 672–677.

[83] C.G. Slatore, S.E. Golden, L. Ganzini, R.S. Wiener and D.H. Au, Distress and patient-centered communication among veterans with incidental (not screen-detected) pulmonary nodules. A cohort study, *Ann Am Thorac Soc* **12** (2015), 184–192.

[84] M.R. Freiman, J.A. Clark, C.G. Slatore, M.K. Gould, S. Woloshin, L.M. Schwartz and R.S. Wiener, Patients' knowledge, beliefs, and distress associated with detection and evaluation of incidental pulmonary nodules for cancer: results from a multicenter survey, *J Thorac Oncol* **11** (2016), 700–708.

[85] A. Verghese, N.H. Shah and R.A. Harrington, What this computer needs is a physician: Humanism and artificial intelligence, *JAMA* **319** (2018), 19–20.

[86] B. Van Calster, L. Wynants, J.F.M. Verbeek, J.Y. Verbakel, E. Christodoulou, A.J. Vickers, M.J. Roobol and E.W. Steyerberg, Reporting and interpreting decision curve analysis: A guide for investigators, *Eur Urol* **74** (2018), 796–804.

[87] A.J. Vickers, A.M. Cronin, E.B. Elkin and M. Gonen, Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers, *BMC Med Inform Decis Mak* **8** (2008), 53.

[88] A.J. Vickers and E.B. Elkin, Decision curve analysis: a novel method for evaluating prediction models, *Med Decis Making* **26** (2006), 565–574.

[89] A.J. Vickers, B. van Calster and E.W. Steyerberg, A simple, step-by-step guide to interpreting decision curve analysis, *Diagn Progn Res* **3** (2019), 18.

[90] M. Fitzgerald, B.R. Saville and R.J. Lewis, Decision curve analysis, *JAMA* **313** (2015), 409–410.

[91] O.Y. Raji, S.W. Duffy, O.F. Agbaje, S.G. Baker, D.C. Christiani, A. Cassidy and J.K. Field, Predictive accuracy of the Liverpool Lung Project risk model for stratifying patients for computed tomography screening for lung cancer: a case-control and cohort validation study, *Ann Intern Med* **157** (2012), 242–250.

[92] M.M. Siddiqui, S. Rais-Bahrami, B. Turkbey, A.K. George, J. Rothwax, N. Shakir, C. Okoro, D. Raskolnikov, H.L. Parnes, W.M. Linehan, M.J. Merino, R.M. Simon, P.L. Choyke, B.J. Wood and P.A. Pinto, Comparison of MR/ultrasound fusion-guided biopsy with ultrasound-guided biopsy for the diagnosis of prostate cancer, *JAMA* **313** (2015), 390–397.

[93] S.G. Baker, Putting risk prediction in perspective: relative utility curves, *J Natl Cancer Inst* **101** (2009), 1538–1542.

[94] S.G. Baker, Decision Curves and Relative Utility Curves, *Med Decis Making* **39** (2019), 489–490.

[95] S.G. Baker and B.S. Kramer, Evaluating prognostic markers using relative utility curves and test tradeoffs, *J Clin Oncol* **33** (2015), 2578–2580.

[96] S.G. Baker, B. Van Calster and E.W. Steyerberg, Evaluating a new marker for risk prediction using the test tradeoff: an update, *Int J Biostat* **8** (2012).

[97] M.N. Kammer, D.J. Rowe, S.A. Deppen, E.L. Grogan, A.M. Kaizer, A.E. Baron and F. Maldonado, The intervention probability curve: modeling the practical application of threshold-guided decision-making, evaluated in lung, prostate, and ovarian cancers, *Cancer Epidemiol Biomarkers Prev* **31** (2022), 1752–1759.

[98] R. Paez, M.N. Kammer, A. Balar, D.A. Lakhani, M. Knight, D. Rowe, D. Xiao, B.E. Heideman, S.L. Antic, H. Chen, S.C. Chen, T. Peikert, K.L. Sandler, B.A. Landman, S.A. Deppen, E.L. Grogan and F. Maldonado, Longitudinal lung cancer prediction convolutional neural network model improves the classification of indeterminate pulmonary nodules, *Sci Rep* **13** (2023), 6157.

[99] M.N. Kammer, D.A. Lakhani, A.B. Balar, S.L. Antic, A.K. Kussrow, R.L. Webster, S. Mahapatra, U. Barad, C. Shah, T. Atwater, B. Diergaarde, J. Qian, A. Kaizer, M. New, E. Hirsch, W.J. Feser, J. Strong, M. Rioth, Y.E. Miller, Y. Balagurunathan, D.J. Rowe, S. Helmey, S.C. Chen, J. Bauza, S.A. Deppen, K. Sandler, F. Maldonado, A. Spira, E. Billatos, M.B. Schabath, R.J. Gillies, D.O. Wilson, R.C. Walker, B. Landman, H. Chen, E.L. Grogan, A.E. Baron, D.J. Bornhop and P.P. Massion, Integrated biomarkers for the management of indeterminate pulmonary nodules, *Am J Respir Crit Care Med* **204** (2021), 1306–1316.

[100] S.H. Park and K. Han, Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction, *Radiology* **286** (2018), 800–809.

[101] E.J. Topol, High-performance medicine: the convergence of human and artificial intelligence, *Nat Med* **25** (2019), 44–56.