Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

LUNGx Challenge for computerized lung nodule classification

Samuel G. Armato, III
Karen Drukker
Feng Li
Lubomir Hadjiiski
Georgia D. Tourassi
Roger M. Engelmann
Maryellen L. Giger
George Redmond
Keyvan Farahani
Justin S. Kirby
Laurence P. Clarke

SPIE.

Samuel G. Armato, , IIIKaren Drukker, Feng Li, Lubomir Hadjiiski, Georgia D. Tourassi, Roger M. Engelmann, Maryellen L. Giger, George Redmond, Keyvan Farahani, Justin S. Kirby, Laurence P. Clarke, "LUNGx Challenge for computerized lung nodule classification," *J. Med. Imag.* **3**(4), 044506 (2016), doi: 10.1117/1.JMI.3.4.044506.

LUNGx Challenge for computerized lung nodule classification

Samuel G. Armato III,[a,*] Karen Drukker,[a] Feng Li,[a] Lubomir Hadjiiski,[b] Georgia D. Tourassi,[c] Roger M. Engelmann,[a] Maryellen L. Giger,[a] George Redmond,[d] Keyvan Farahani,[d] Justin S. Kirby,[e] and Laurence P. Clarke[d]

[a]The University of Chicago, Department of Radiology, 5841 South Maryland Avenue, MC 2026, Chicago, Illinois 60637, United States
[b]University of Michigan, Department of Radiology, 1500 East Medical Center Drive, Ann Arbor, Michigan 48109, United States
[c]Health Data Sciences Institute, Biomedical Science and Engineering Center, Oak Ridge National Laboratory, P.O. Box 2008 MS6085 Oak Ridge, Tennessee 37831-6085, United States
[d]National Cancer Institute, Cancer Imaging Program, Division of Cancer Treatment and Diagnosis, 9609 Medical Center Drive, Bethesda, Maryland 20892, United States
[e]Leidos Biomedical Research, Inc., Frederick National Laboratory for Cancer Research, Cancer Imaging Program, 8560 Progress Drive, Frederick, Maryland 21702, United States

Abstract. The purpose of this work is to describe the LUNGx Challenge for the computerized classification of lung nodules on diagnostic computed tomography (CT) scans as benign or malignant and report the performance of participants' computerized methods along with that of six radiologists who participated in an observer study performing the same Challenge task on the same dataset. The Challenge provided sets of calibration and testing scans, established a performance assessment process, and created an infrastructure for case dissemination and result submission. Ten groups applied their own methods to 73 lung nodules (37 benign and 36 malignant) that were selected to achieve approximate size matching between the two cohorts. Area under the receiver operating characteristic curve (AUC) values for these methods ranged from 0.50 to 0.68; only three methods performed statistically better than random guessing. The radiologists' AUC values ranged from 0.70 to 0.85; three radiologists performed statistically better than the best-performing computer method. The LUNGx Challenge compared the performance of computerized methods in the task of differentiating benign from malignant lung nodules on CT scans, placed in the context of the performance of radiologists on the same task. The continued public availability of the Challenge cases will provide a valuable resource for the medical imaging research community. [DOI: 10.1117/1.JMI.3.4.044506]

Keywords: lung nodule; classification; computer-aided diagnosis; challenge; computed tomography; image analysis.

Paper 16082R received May 22, 2016; accepted for publication Nov. 17, 2016; published online Dec. 19, 2016.

1 Introduction

The computer-aided diagnosis (CAD) research community, and more recently the radiomics community, is actively engaged in the development of computerized methods for a wide variety of applications in radiology. Medical imaging challenges allow for a direct comparison of algorithms in that all algorithms abide by the same rules on the same images evaluated in the same manner. The conduct of "grand challenges" helps to shape the landscape of CAD research by encouraging development of methodologies for targeted tasks with unique or timely clinical relevance. Grand challenges also foster the concept of "open science" and provide the resources necessary for friendly competition among research groups with the overall goal of advancing the field by generating interest and collaborations among investigators.

The literature contains many publications from various research groups on the performance of their respective CAD systems applied to their own datasets; however, the impact on system performance of well-known factors such as database composition, "truth" definition, and scoring metric[1–3] confounds the ability to evaluate the true forward momentum of the field based on such varied presentations. Challenges normalize for these factors by directly comparing different algorithms designed for a specific radiologic task performed on a common set of images. Within the documented parameters of a challenge, all participating algorithms must adhere to the same rules and are evaluated according to an established performance assessment paradigm. Challenges present an opportunity to explore similarities and differences in the ability of different algorithms to perform the same task under the same conditions.

The thoracic imaging research community has hosted a number of successful challenges that span a range of tasks,[4,5] including lung nodule detection,[6] lung nodule change, vessel segmentation,[7] and vessel tree extraction.[8] The recent LUNGx Challenge involved computerized classification of lung nodules as benign or malignant on diagnostic computed tomography (CT) scans.[9] The LUNGx Challenge (hereafter referred to as "the Challenge") was a collaborative effort sponsored by SPIE and the American Association of Physicists in Medicine (AAPM) along with the National Cancer Institute (NCI). The Challenge was conducted during the months leading up to the 2015 SPIE Medical Imaging Symposium, with results presented at a special panel session within the CAD Conference during the symposium. Lessons learned from this Challenge

*Address all correspondence to: Samuel G. Armato III, E-mail: s-armato@uchicago.edu

were reported previously through an editorial.[9] The purpose of this paper is to summarize the results obtained by groups that participated in the Challenge and to report the diagnostic performance of experienced radiologists in an observer study using the same set of CT scans.

2 Methods

2.1 *Dataset*

All thoracic CT scans used in the Challenge were obtained retrospectively from the clinical archive at The University of Chicago with approval from the local Institutional Review Board. A deidentification process[10] removed all protected health information from the DICOM headers of the images (other information within the DICOM headers remained intact) prior to upload to the NCI's publicly accessible The Cancer Imaging Archive (TCIA),[11] which hosted the images and associated data for the Challenge.[12] Each scan used in the Challenge consisted of a single transaxial series with full thoracic coverage from unique patients. All scans had been acquired on Philips Brilliance scanners with a "D" (overenhancing) convolution kernel and a 1.0-mm slice thickness. All nodules within these scans were determined by a radiologist (FL) to be either primary lung cancer or benign based on follow-up imaging and/or pathologic assessment. It is important to note that the Challenge presented a classification task not a detection task.

Groups participating in the Challenge were expected to have an already developed and trained system for the computerized classification of lung nodules. Accordingly, a set of 10 calibration scans (all but two were contrast enhanced) was made available to participating groups rather than a complete set of training scans. These 10 calibration scans (five males, five females; median age: 65 years) were intended to assist groups evaluate the compatibility of the Challenge scans with their own algorithms (in terms of image acquisition parameters and DICOM file structure) and were not intended for algorithm development or classifier training. The spatial coordinates of the approximate center of each nodule were provided along with the diagnosis (benign or malignant) for each nodule in the calibration scans. Five of the 10 calibration scans contained a single confirmed benign nodule (two confirmed based on nodule stability for at least 2 years, two confirmed based on nodule resolution, and one confirmed based on pathologic assessment), and the other five scans contained a single pathology-confirmed malignant nodule (two small cell carcinomas, one poorly differentiated adenocarcinoma, one moderately differentiated adenocarcinoma, and one nonsmall cell carcinoma). The Challenge announcement explicitly instructed participants to consider the calibration set as representative of the technical aspects of the test set and not necessarily as representative of the lung nodule variety and difficulty level to be expected in the test set.

Approximately 7 weeks after the calibration scans became available, the test set of 60 scans with a total of 73 nodules (13 scans contained two nodules each) was released. The 60 test scans (23 males, 37 females; median age: 60.5 years) contained 37 benign nodules (including 13 confirmed based on nodule stability for at least 2 years, 19 confirmed based on nodule resolution, and five confirmed based on pathologic assessment) and 36 malignant nodules (including 15 adenocarcinomas, nine nonsmall cell carcinomas, seven small cell carcinomas, two carcinoid tumors, one squamous cell carcinoma, and two nodules suspicious for malignancy). One scan contained both a malignant nodule (nonsmall cell carcinoma) and a benign nodule. Considering the 30 scans with malignant nodules, there were 11 males and 19 females, with a median age of 61 years (range: 44 to 86 years); considering the 31 scans with benign nodules, there were 12 males and 19 females, with a median age of 61 years (range: 18 to 79 years).

Nodule size was measured by a radiologist (FL) using electronic calipers as the longest nodule diameter according to the Response Evaluation Criteria in Solid Tumors (RECIST) guidelines.[13] Scans in the test set were deliberately selected to achieve some level of nodule size matching based on simplistic visual inspection of nodule size histograms (although this fact was not disclosed to participants), since nodule size is a well-known predictor of malignancy.[14] The 37 benign nodules had a mean RECIST size of 15.8 mm (standard deviation: 8.2 mm), and the 36 malignant nodules had a mean RECIST size of 18.6 mm (standard deviation: 6.7 mm); a Student's *t*-test for the differences in means yielded a *p*-value of 0.12. Nodule size and other nodule characteristics are reported in Table 1. The spatial coordinates of the approximate center of each nodule were provided for each nodule in the test scans; nodule diagnosis and nodule size were not released to participants.

Table 1 Characteristics of the 73 nodules in the LUNGx Challenge test set.

	Malignant nodules ($n = 36$)	Benign nodules ($n = 37$)	*p*-value
RECIST-based nodule size (mm)			
Mean (standard deviation)	18.6 (6.7)	15.8 (8.2)	0.12[a]
Median (range)	13.9 (5.7 to 45.0)	17.1 (4.6 to 34.6)	
Nodule solidity			
Nonsolid ($n = 4$)	2 (50%)	2 (50%)	
Part solid ($n = 10$)	5 (50%)	5 (50%)	
Solid ($n = 59$)	29 (49%)	30 (51%)	0.99[b]
Nodule location			
Left lower lobe ($n = 12$)	5 (42%)	7 (58%)	
Left upper lobe ($n = 23$)	13 (57%)	10 (43%)	
Right lower lobe ($n = 15$)	5 (33%)	10 (67%)	
Right middle lobe ($n = 6$)	2 (33%)	4 (67%)	
Right upper lobe ($n = 17$)	11 (65%)	6 (35%)	0.34[b]
Spiculation			
Absent ($n = 43$)	17 (40%)	26 (60%)	
Present ($n = 30$)	19 (63%)	11 (37%)	0.045[b]

Note: RECIST = Response Evaluation Criteria in Solid Tumors.
[a]*p*-value computed from Student's *t*-test.
[b]*p*-value computed from chi-square test.

Among the 60 test set scans, 29 had been acquired on Brilliance 16 or 16P scanners, and 31 had been acquired on Brilliance 64 scanners. The tube peak potential energy used for scan acquisition was 120 kV ($n = 53$) or 140 kV ($n = 7$). Tube current ranged from 240 to 500 mA (mean: 410.1 mA). The in-plane pixel size ranged from 0.549 to 0.900 mm (mean: 0.685 mm). All images had a matrix size of 512×512 pixels. All but 13 test set scans were contrast enhanced.

2.2 *Challenge Task*

Participants were expected to independently apply their nodule classification algorithms to the 73 lung nodules in the test set and assign a "score" to each nodule on a continuous scale as a metric for the likelihood of nodule malignancy. Participants also reported whether low or high scores were expected to represent malignant nodules for the output of their method. Each participating group was required to send in an e-mail to the Challenge organizers a single file with the 73 scores output by their method by no later than 25 days after the test set had become available.

The malignancy rating output from 15 separate methods was submitted by 12 different groups (the participants). Four methods were withdrawn from the Challenge: one participant withdrew its two methods because the area under the receiver operating characteristic curve (AUC) values they achieved were not high enough to meet the expectations of that participant, and two other methods were withdrawn when it became known that these participants had obtained malignancy ratings for the test set nodules from local radiologists to train their methods (thus negating the independence of the test set for these two methods). The final contributions to the Challenge then became 11 separate methods from 10 different participants.

A majority of the computerized methods followed the same general approach: (1) nodule segmentation, (2) feature extraction, (3) feature selection, and (4) classifier application (Table 2). Among these methods, nodule segmentation techniques included gray-level thresholding, morphological operations, region growing, and graph-cut-based optimal surface detection. The spatial location of the nodule center provided to participants often served as the seed point for nodule segmentation, although one method required construction of a coarse manual outline around the nodule to initialize segmentation. Extracted features spanned first-order gray-level features, texture features, and shape-based features. Feature selection techniques included principal component analysis, mutual information, and a random-forest-based approach. Classifiers included support vector machine, support vector regressor, random forest, and WEKA. In addition, one method used radiologist-defined semantic labels to capture nodule characteristics and form a discriminant function, one method applied deep learning and used a convolutional neural network to classify image regions of interest, and another method used the frequency of histogram-equalized pixel values within a nodule and its local neighborhood to define a rule-based malignancy score. A final method attempted to quantify nodule perfusion as a surrogate for malignancy based on contrast inflow patterns from the surrounding vasculature; region growing was performed outward from the specified nodule centroid and inward along intersecting vessels, and ratios of pixel values from both directions at each iteration of this bidirectional region growing were used to quantify perfusion. Participants trained their systems on the LUNGx calibration scans, cases from the National Lung Screening Trial,[15] the Lung Image Database Consortium Database,[16] or local collections of cases.

2.3 *Performance Evaluation*

With internal knowledge of the actual nodule diagnoses, the organizers used receiver operating characteristic (ROC) analysis[17,18] to assess, for each participant's method, classification performance in the task of distinguishing between malignant and benign lung nodules. The AUC served as the performance metric and was provided to each participant for their submitted results after being computed by the organizers. The method with the highest AUC value was selected as the winner of the Challenge; a statistically significant difference in performance with respect to the method with the next highest AUC value was not required.

The performance of each individual method was compared (1) with random guessing (AUC = 0.5) and (2) with the classification performance achieved when using just the attribute of RECIST-based nodule size as determined by the study radiologist. For these comparisons, bootstrapping was used with 1000 iterations to assess statistical significance (using two-sided 95% confidence intervals for the difference in AUC and associated p-values) and noninferiority (lower-bound of the single-sided 90% confidence intervals for the difference in AUC). The Holm–Bonferroni correction for multiple comparisons was used.[19]

2.4 *Observer Study*

After the Challenge was completed, an observer study was conducted to provide real-world context for the performance levels of the various computer algorithms within the Challenge. Six attending thoracic radiologists (median years of experience 14 years; range: 7 to 41 years) independently (and completely manually without any computer diagnostic aid) performed the same classification task on the same 73 lung nodules from the Challenge test set. An interface was developed that allowed a user to raster through all section images of a CT scan, manipulate the visualization settings (window, level, and zoom), and view relevant information from the image DICOM headers (patient gender, patient age, and image reconstruction kernel) (Fig. 1). The interface allowed a user to directly view the nodule location within the images based on the same spatial coordinates for the nodule center that were provided to Challenge participants. The radiologists used a slider bar on the interface to mark their assessment of nodule malignancy status, with one end of the continuous scale labeled "definitely benign" and the other end of the scale labeled "definitely malignant." After being shown the 10 calibration scans (and the nodule diagnoses) to become familiar with the interface and understand the classification task, each radiologist was shown, in random order, each of the 73 nodules from the Challenge test set; the interface captured their malignancy ratings as converted from the marked slider bar location to a continuous 0-1 scale.

The observer study was designed to replicate as closely as possible a manual implementation of the actual Challenge, with the same information provided and the same constraints imposed. It is important to note that the radiologists only had access to the images and relevant metadata available in the DICOM header of the images; other patient information commonly used in clinical practice was not available. The only

Table 2 AUC values for the 11 computerized methods and six observers in the task of differentiating between malignant and benign lung nodules.

Method	AUC value	SE	Nodule segmentation	Classifier	Cases to train
1	0.50	0.068	Voxel-intensity-based segmentation	SVM	LUNGx calibration
2	0.50	0.056	Region growing	WEKA	NLST
3	0.54	0.067	None required	Rules based on histogram-equalized pixel frequencies	LUNGx calibration
4	0.54	0.066	Bidirectional region growing	Uses tumor perfusion surrogate	LUNGx calibration
5	0.55	0.067	Region growing	WEKA	NLST
6	0.56	0.054	Graph-cut-based surface detection	Random forest	LIDC
7	0.59	0.066	Manual initialization, gray-level thresholding, morphological operations	SVM	LUNGx calibration
8	0.59	0.053	None required	Convolutional neural network	LIDC
9	0.61	0.054	GrowCut region growing with automated initial label points	SVM	NLST
10	0.66	0.063	Radiologist-provided nodule semantic ratings	Discriminant function	LUNGx calibration
11	0.68	0.062	Semiautomated thresholding	Support vector regressor	In-house dataset
Observer					
1	0.70	0.060			
2	0.75	0.057			
3	0.78	0.046			
4	0.82	0.049			
5	0.83	0.047			
6	0.85	0.044			

potentially biasing instruction given to the radiologists was to encourage them to use the entire range of the rating scale as selecting extreme ends of the scale would indicate that the radiologist could not possibly be more confident about the diagnosis. The organizers applied ROC analysis to evaluate the nodule classification performance of each radiologist.

3 Results

3.1 *Participating Computerized Methods*

ROC curves for the 11 classification methods are shown in Fig. 2. The AUC values ranged from 0.50 to 0.68 (with standard error 0.06 for the latter) (Table 2). Only three of the methods performed better than random guessing, with p-values of 0.006, 0.008, and 0.048; these p-values do not remain statistically significant after the Holm–Bonferroni correction. For comparison, radiologist-determined nodule size (per RECIST) achieved an AUC of 0.64 (standard error 0.06), which alone was significantly better than random guessing (p-value 0.034). Given the statistical power of the Challenge, none of the computerized methods achieved a level of performance that was significantly better than RECIST nodule size alone (even for what was intended to be a size-matched dataset) ($p > 0.05$), and noninferiority to nodule size alone could not be established with, for the best-performing method, a lower bound of the 90% confidence interval for the difference in AUC values of -0.054.

3.2 *Observer Study*

ROC curves for the six radiologists who participated in the observer study (no computer aid was used) are shown in Fig. 3. The AUC values ranged from 0.70 to 0.85 with standard errors ranging from 0.04 to 0.06 (Table 2). The mean AUC value across all six radiologists was 0.79 (standard deviation 0.06). Three radiologists performed better than the best-performing computer method, with p-values of 0.023, 0.027, and 0.047; the performance of the other three radiologists failed to reach statistical significance relative to the best-performing computer

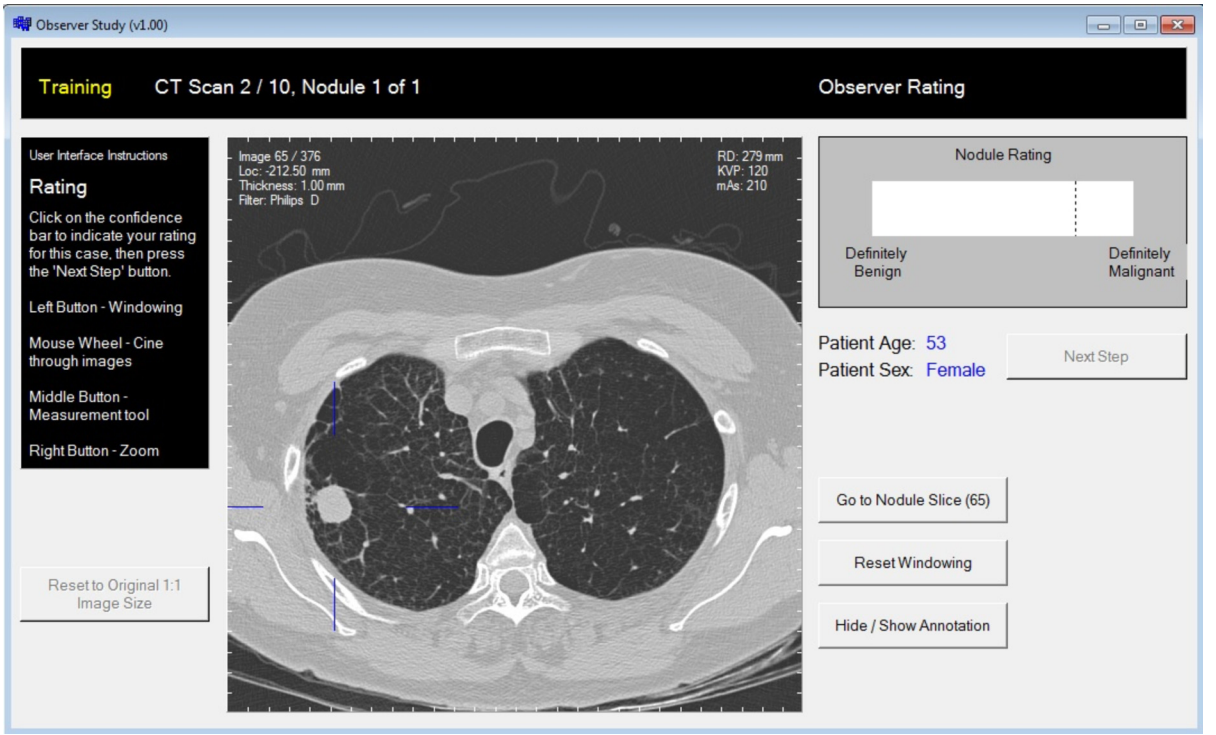

Fig. 1 The interface developed for the observer study allowed a user to raster through all section images of a scan, manipulate the visualization settings, and view relevant patient and image-acquisition information from the image DICOM headers. Nodules for evaluation were demarcated with blue crosshairs. Radiologists used the slider bar to mark their assessment of nodule malignancy.

method ($p = 0.140$, 0.242, and 0.773). These p-values are not corrected for multiple comparisons; note that when correcting for multiple comparisons using Holm–Bonferroni, all of these differences fail to achieve statistical significance. Figure 4 shows examples of benign and malignant nodules that received discordant malignancy ratings between the computer methods and the radiologists.

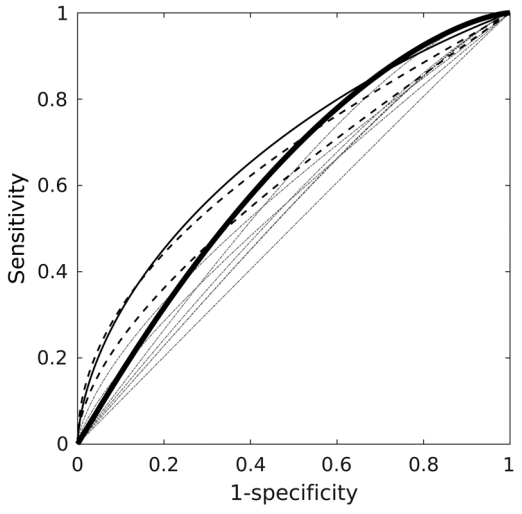

Fig. 2 ROC curves for the 11 participating classification methods, with AUC values ranging from 0.50 to 0.68. The thick solid curve is for radiologist-determined nodule size alone (AUC = 0.62). The two dashed curves outperformed random guessing but failed to be statistically different from nodule size. The thin solid curve is for the winning algorithm, which outperformed random guessing and was noninferior to nodule size alone.

4 Discussion

Challenges should be approached by both organizers and participants as a friendly competition within the research community, designed to foster interest in the designated task and encourage innovation in the field. Challenge organizers have a responsibility to craft a set of rules that are as complete, clear, and logical as possible, while anticipating possible misinterpretations or confusion on the part of participants.

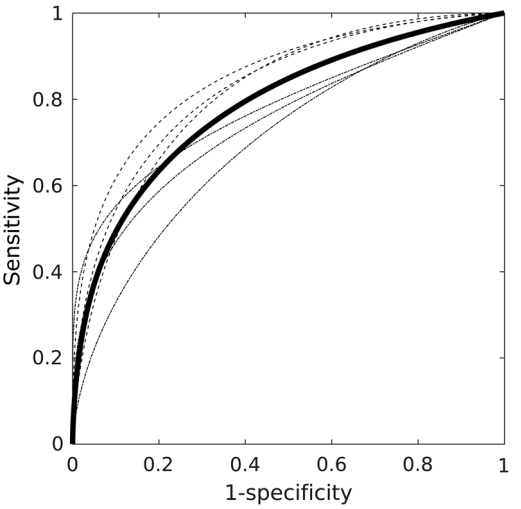

Fig. 3 ROC curves for the six radiologists from the observer study. The thick solid curve is for the radiologists as a group. The dashed curves represent those radiologists who significantly outperformed the CAD winner. The AUC values ranged from 0.70 to 0.85, with a mean AUC value across all six radiologists of 0.79.

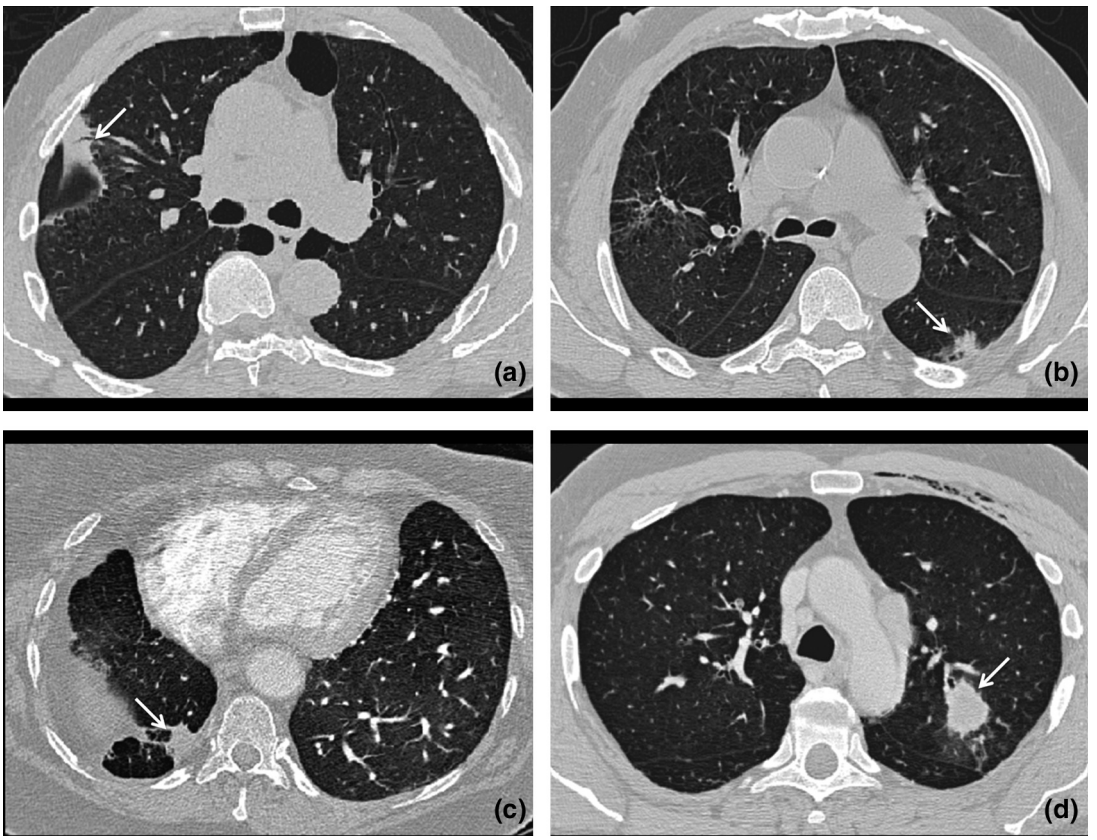

Fig. 4 (a) A benign nodule (arrow) for which the best-performing method returned (correctly) a low likelihood of malignancy score but to which all radiologists assigned higher malignancy ratings. (b) A malignant nodule (arrow) for which the best-performing method returned (correctly) a high likelihood of malignancy score but to which all radiologists assigned lower malignancy ratings. (c) A benign nodule (arrow) that was misdiagnosed by the best-performing method but that received a low malignancy rating from the best-performing radiologist. (d) A malignant nodule (arrow) that was misdiagnosed by the best-performing method but that received a high malignancy rating from the best-performing radiologist.

Organizers have a further responsibility to collect a well-vetted set of relevant cases, to establish a performance assessment process for the evaluation of results, and to create an infrastructure for case dissemination, communication, and result submission. Challenge participants have an obligation to abide by the rules, to approach the challenge with commitment and scientific rigor, and to accept the final performance analysis. These concepts, along with other lessons learned from the LUNGx Challenge, were reported previously in a *Journal of Medical Imaging* guest editorial.[9]

The Challenge provided a common dataset of diagnostic thoracic CT scans with lung nodules that had confirmed malignancy status. While such a common dataset is essential to a challenge, it is important to note that the performance of the participating methods was assessed only for this single collection of images acquired with a rather homogeneous set of technical parameters from one institution; extrapolation of the Challenge results achieved by individual methods to a general setting or to another cohort of CT scans would be flawed and misleading.

Participants deserve credit for their willingness to apply their algorithms to images acquired with parameters that might differ substantially from those under which their algorithms had been developed, especially since this challenge did not provide a proper set of training scans. Although generalizability is a desired goal for mature CAD methodologies, it is impractical to expect a method to achieve this goal without some level of comprehensive training. The Challenge could have acknowledged this issue by providing either a more extensive set of training scans or a set of test scans with greater heterogeneity in imaging parameters from multiple institutions; however, the additional case-collection effort would have been burdensome.

The topic of the LUNGx Challenge is especially timely given the recent development of lung nodule risk models and attention toward the clinical deployment of such models,[20–24] which benefit from the inclusion of imaging findings. Cancer risk models, however, require clinical and demographic information beyond that provided by the LUNGx Challenge. Rather than seeking to promote the development of image-based risk models, the Challenge was designed to advance the efforts of the CAD community to develop lung nodule classification methods that differentiate between benign and malignant nodules from a single-timepoint CT scan.[25–39]

Any observer study necessarily forces radiologists to perform a clinical task under conditions that differ from their normal mode of decision making. Patient demographics (except for gender and age), clinical history, and prior imaging studies were not available during the observer study, because this information was not available to Challenge participants despite the fact that some participants (and all the radiologists) might have reasonably desired information such as smoking history and nodule size from a previous CT scan. Radiologists, however, could

exploit additional cues that even the most advanced computerized algorithms currently lack the sophistication to incorporate. For example, although the entire CT scan was available both to participants during the Challenge and to radiologists during the observer study, the presence of concomitant disease (e.g., emphysema or pleural plaques) would influence a radiologist's assessment of nodule malignancy but would be ignored by lung nodule classification algorithms; advanced CAD systems to evaluate the presence of emphysema or pleural disease have been reported, but the merging of information derived from such methods with algorithms to compute the probability of lung nodule malignancy has not been achieved. A dataset of CT scans demonstrating only a single lung nodule (some scans in the Challenge test set contained two nodules) with no other abnormality could have been collected, but such a restriction would have greatly increased the case-collection effort, reduced the number of available qualifying scans (with a reduction in statistical power), and further skewed the case distribution away from clinical reality.

As another example of the potential for human insight to impact an observer study, radiologists familiar with observer studies will reasonably suspect that the test set of scans has been enriched in an attempt to achieve a roughly equal number of benign and malignant nodules with matched size characteristics. Therefore, larger nodules that would typically be viewed more suspiciously in a clinical setting might be more easily assessed as benign in the context of an observer study. Interestingly, despite a specific (although subjective) effort to achieve size matching between the benign and malignant nodules, nodule size alone attained a level of nodule classification performance that was statistically better than random guessing; future efforts to attain matching of characteristics between dataset classes should confirm the success of that matching through application of the performance-assessment paradigm to each such characteristic alone.

Nodule size matching was important to the design of this challenge, although size matching certainly would be detrimental to other tasks. Cases with a clinically relevant distribution of nodule size (as well as other nodule characteristics) would be critical, for example, in efforts to develop a model to predict cancer likelihood; other information not provided in the LUNGx Challenge, such as smoking history, would also be essential to such a task, and a collection of cases obtained from more than just a single institution would be required. The purpose of the Challenge, however, was to use image-based nodule features (other than simply size alone) to differentiate benign from malignant nodules. It is standard practice in the CAD community to use "enriched" datasets with case distributions that do not follow clinical experience and datasets that have been intentionally matched between two groups of cases (e.g., benign cases and malignant cases or gene mutation carriers and nonmutation) for certain attributes such as patient age, patient gender, and lesion size. In effect, attribute matching allows a study to control for "easier" factors so that the study may focus on more complex aspects of the given task. The LUNGx Challenge dataset was both enriched (with respect to the proportion of benign and malignant nodules) and matched (with respect to nodule size between the benign and malignant nodules); neither of these points was disclosed to participants, otherwise this information could have been used to bias the performance of the computerized systems. It should again be noted that despite the subjective attempt to achieve size matching between the benign and malignant nodules, nodule size alone attained a level of nodule classification performance that was statistically better than random guessing.

The rules of a challenge and the expectations of participants must be clearly conveyed from the beginning, otherwise a wide variety of interpretations may be possible. For example, the 10 calibration scans were intended only to assist groups assess the compatibility of Challenge cases with their own algorithms; nevertheless, some groups attempted to use this intentionally small set of calibration scans for algorithm development and/or classifier training. As another example, despite the implicit assumption that the participating methods would be fully automated with no human involvement, some groups incorporated input extracted by local radiologists from the Challenge test scans, including manual nodule outlines, semantic labels that captured nodule characteristics, and nodule malignancy ratings; the first two of these radiologist inputs were subsequently determined by the organizers to be acceptable (although unexpected), while the method that used radiologist malignancy ratings was one of the withdrawn methods.

Case collection for challenges will almost always be a laborious task, and the more selective the criteria imposed on the desired cases or the more positioned the cases are for subsequent subset analyses, the greater this burden becomes. A challenge, for example, could benefit from the collection of cases that are uniquely complex for the designated radiologic task, but such cases are less common; often organizers must forego attempts to obtain targeted cases for a more practical collection of more routine cases from the clinical archives of one or multiple institutions. The LUNGx Challenge collected cases from a defined time period containing "confirmed" (a term with several possible definitions) malignant and benign lung nodules, with a subjective sense of size matching imposed as the only constraint; any attempts at subset analysis (e.g., performance evaluation based on malignant nodule histology) would have been statistically meaningless, since the cases were not collected to achieve sufficient numbers across possible relevant subsets (and case collection would have been much more time consuming to meet such a requirement).

The continued public availability of the LUNGx Challenge cases and associated data, including the newly released diagnosis (truth) information for the test cases (available for download from Ref. 12), provides a valuable resource for the medical imaging research community. Anyone wishing to use these data for presentation or publication purposes should acknowledge the SPIE, the AAPM, the NCI, and The University of Chicago. It becomes the responsibility of the users of these data to clearly report the manner in which they developed, trained, and evaluated their systems.

5 Conclusion

The LUNGx Challenge was a successful scientific challenge for the computerized classification of lung nodules on CT scans jointly sponsored by the SPIE, AAPM, and NCI. Ten participating groups from academia and industry applied 11 computerized methods to the 73 lung nodules in the test set of scans; these methods ranged from fully automated to semiautomated with varying levels of radiologist input. Only three of these methods performed better than random guessing within the statistical limits of the Challenge. To place the performance of the computerized methods into a real-world context, an observer study was conducted with six attending radiologists manually performing

the same Challenge task. Three of the radiologists performed better than the best-performing computer method. Challenges should be approached by both organizers and participants as a friendly competition within the research community, designed to foster interest in the designated task and encourage innovation in the field. The continued public availability of the Challenge cases will provide a valuable resource for the medical imaging research community into the future.

Disclosures

SGA, KD, MLG, and RE receive royalties and licensing fees through The University of Chicago related to computer-aided diagnosis.

Acknowledgments

This paper is dedicated to the memory of Laurence P. Clarke, PhD, whose vision helped shape not only the LUNGx Challenge but also, more importantly, the broader imaging research landscape of open science, consensus building, and collegiality that provides context for the significance of grand challenges. He worked tirelessly to advance research objectives by helping to surmount obstacles, building bridges across institutions or governmental agencies, and selflessly advocating on behalf of other investigators who shared his devotion to imaging research. Dr. Clarke's career had a dramatic impact on the current state of medical imaging research around the world. His leadership was profound, his scientific contributions substantial, his intentions sincere, his humor infectious, and his personality genuine; he will be greatly missed by all his coauthors, colleagues, and friends. The authors would like to express their appreciation to Angela Keyser and Shayna Knazik from the AAPM and Diane Cline and Sandy Hoelterhoff from the SPIE for their assistance in making this challenge a reality. The LUNGx Challenge was supported by SPIE, AAPM, and NCI. The authors are grateful to the groups that participated in the LUNGx Challenge, including groups led by Yoganand Balagurunathan (H. Lee Moffitt Cancer Center), Wookjin Choi (University of Maryland School of Medicine, now at Memorial Sloan Kettering Cancer Center), Samuel Hawkins (University of South Florida), Xiaojie Huang (GE Global Research), Lyndsey Pickup (Mirada Medical, now at Optellum), Jiangdian Song (Northeastern University, China), Yubing Tong (University of Pennsylvania), and Rafael Wiemker (Philips Research Labs Hamburg). These participating groups deserve recognition for their work to develop their lung nodule classification methods and for their contributions to the success of the Challenge. The authors would like to thank the radiologists who participated in the observer study: Prachi Agarwal, MD, Aamer Chughtai, MD, Jonathan Chung, MD, Heber MacMahon, MD, Steven Montner, MD, and Smita Patel, MD. This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. This manuscript has been authored in part by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The US Government retains and the publisher, by accepting the article for publication, acknowledges that the US Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

References

1. R. M. Nishikawa et al., "Effect of case selection on the performance of computer-aided detection schemes," *Med. Phys.* **21**, 265–269 (1994).
2. R. M. Nishikawa et al., "Variations in measured performance of CAD schemes due to database composition and scoring protocol," *Proc. SPIE* **3338**, 840–844 (1998).
3. G. Revesz, H. L. Kundel, and M. Bonitatibus, "The effect of verification on the assessment of imaging techniques," *Invest. Radiol.* **18**, 194–198 (1983).
4. B. van Ginneken, "Why Challenges?" http://grand-challenge.org/Why_Challenges/ (December 2015).
5. K. Murphy, "Development and evaluation of automated image analysis techniques in thoracic CT," PhD Thesis, Utrecht University (2011).
6. B. van Ginneken et al., "Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study," *Med. Image Anal.* **14**, 707–722 (2010).
7. R. D. Rudyanto et al., "Comparing algorithms for automated vessel segmentation in computed tomography scans of the lung: the VESSEL12 study," *Med. Image Anal.* **18**, 1217–1232 (2014).
8. P. Lo et al., "Extraction of airways from CT (EXACT'09)," *IEEE Trans. Med. Imaging* **31**, 2093–2107 (2012).
9. S. G. Armato, III et al., "The LUNGx Challenge for computerized lung nodule classification: reflections and lessons learned," *J. Med. Imaging* **2**, 020103 (2015).
10. S. G. Armato, III et al., "Research imaging in an academic medical center," *Acad. Radiol.* **19**, 762–771 (2012).
11. K. Clark et al., "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository," *J. Digit Imaging* **26**, 1045–1057 (2013).
12. S. G. Armato, III et al., "SPIE-AAPM-NCI lung nodule classification challenge dataset," *Cancer Imaging Arch.* (2015).
13. P. Therasse et al., "New guidelines to evaluate the response to treatment in solid tumors," *J. Nat. Cancer Inst.* **92**, 205–216 (2000).
14. A. C. Jirapatnakul et al., "Characterization of pulmonary nodules: effects of size and feature type on reported performance," *Proc. SPIE* **6915**, 69151E (2008).
15. P. K. Commean et al., "A query tool for investigator access to the data and images of the national lung screening trial," *J. Digital Imaging* **28**, 439–447 (2015).
16. S. G. Armato, III et al., "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans," *Med. Phys.* **38**, 915–931 (2011).
17. L. L. Pesce et al., "Reliable and computationally efficient maximum-likelihood estimation of "proper" binormal ROC curves," *Acad. Radiol.* **14**, 814–829 (2007).
18. C. E. Metz, "ROC methodology in radiologic imaging," *Invest. Radiol.* **21**, 720–733 (1986).
19. S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Stat.* **6**, 65–70 (1979).
20. A. McWilliams et al., "Probability of cancer in pulmonary nodules detected on first screening CT," *N. Engl. J. Med.* **369**, 910–919 (2013).
21. D. Manos et al., "The lung reporting and data system (LU-RADS): a proposal for computed tomography screening," *Can. Assoc. Radiol. J.* **65**, 121–134 (2014).
22. Y. Li and J. Wang, "A mathematical model for predicting malignancy of solitary pulmonary nodules," *World J. Surg.* **36**:830–835 (2012).
23. F. Xiao et al., "Novel and convenient method to evaluate the character of solitary pulmonary nodule-comparison of three mathematical prediction models and further stratification of risk factors," *PLoS One* **8**, e78271 (2013).
24. S. Perandini et al., "Solid pulmonary nodule risk assessment and decision analysis: comparison of four prediction models in 285 cases," *Eur. Radiol.* **26**, 3071–3076 (2016).

25. M. F. McNitt-Gray et al., "A pattern classification approach to characterizing solitary pulmonary nodules imaged on high resolution CT: preliminary results," *Med. Phys.* **26**, 880–888 (1999).
26. S. G. Armato, III et al., "Automated lung nodule classification following automated nodule detection on CT: a serial approach," *Med. Phys.* **30**, 1188–1197 (2003).
27. S. K. Shah et al., "Computer aided characterization of the solitary pulmonary nodule using volumetric and contrast enhancement features," *Acad. Radiol.* **12**, 1310–1319 (2005).
28. T. W. Way et al., "Computer-aided diagnosis of pulmonary nodules on CT scans: segmentation and classification using 3D active contours," *Med. Phys.* **33**, 2323–2337 (2006).
29. H. Chen et al., "Neural network ensemble-based computer-aided diagnosis for differentiation of lung nodules on CT images: clinical evaluation," *Acad. Radiol.* **17**, 595–602 (2010).
30. Y. Zhu et al., "Feature selection and performance evaluation of support vector machine (SVM)-based classifier for differentiating benign and malignant pulmonary nodules by computed tomography," *J. Digital Imaging* **23**, 51–65 (2010).
31. S. K. Dilger et al., "Improved pulmonary nodule classification utilizing quantitative lung parenchyma features," *J. Med. Imaging* **2**, 041004 (2015).
32. K. L. Hua et al., "Computer-aided classification of lung nodules on computed tomography images via deep learning technique," *OncoTargets Ther.* **8**, 2015–2022 (2015).
33. C. Jacobs et al., "Solid, part-solid, or non-solid? Classification of pulmonary nodules in low-dose chest computed tomography by a computer-aided diagnosis system," *Invest. Radiol.* **50**, 168–173 (2015).
34. G. A. Soardi et al., "Assessing probability of malignancy in solid solitary pulmonary nodules with a new Bayesian calculator: improving diagnostic accuracy by means of expanded and updated features," *Eur. Radiol.* **25**, 155–162 (2015).
35. W. Choi and T.-S. Choi, "Genetic programming-based feature transform and classification for the automatic detection of pulmonary nodules on computed tomography images," *Inf. Sci.* **212**, 57–78 (2012).
36. J. Song et al., "Lung lesion extraction using a toboggan based growing automatic segmentation approach," *IEEE Trans. Med. Imaging* **35**, 337–353 (2016).
37. Y. Liu et al., "Quantitative imaging features to predict cancer status in lung nodules," *Proc. SPIE* **9787**, 97870L (2016).
38. S. Hawkins et al., "Predicting malignant nodules from screening CT scans," *J. Thorac. Oncol.* **11**, 2120–2128 (2016).
39. Y. Liu et al., "Radiological image traits predictive of cancer status in pulmonary nodules," *Clin. Can. Res.* (2016) (in press).

Samuel G. Armato III is an associate professor of radiology and the Committee on Medical Physics at The University of Chicago. His research interests involve the development of computer-aided diagnostic (CAD) methods for thoracic imaging, including automated lung nodule detection and analysis in CT scans, semiautomated mesothelioma tumor response assessment, image-based techniques for the assessment of radiotherapy-induced normal tissue complications, and the automated detection of pathologic change in temporal subtraction images.

Karen Drukker has been actively involved in CAD/radiomics research at The University of Chicago for over a decade. Her work has focused on multimodality detection/diagnosis/prognosis of breast cancer and on the performance evaluation of radiomics methods.

Feng Li is a staff scientist in the Department of Radiology and a research radiologist in The Human Imaging Research Office at The University of Chicago. Her research interests include the detection of early lung cancers, analysis of radiologist-missed cancers, classification of malignant and benign lung nodules in chest CT scans or chest radiography, CAD, advanced image techniques, and tumor response assessment.

Lubomir Hadjiiski is a professor in the Department of Radiology at the University of Michigan. He has authored or coauthored more than 115 publications in peer-reviewed journals. His research interests include CAD, neural networks, predictive models, image processing, medical imaging, and control systems. His current research involves design of decision support systems for detection and diagnosis of cancer in different organs and quantitative analysis of image biomarkers for treatment response monitoring.

Georgia D. Tourassi is the director of the Biomedical Science and Engineering Center and the Health Data Sciences Institute at Oak Ridge National Laboratory. She holds a PhD in biomedical engineering from Duke University. Her research interests include biomedical informatics, medical imaging, and computer-aided decision support. She has authored over 200 peer-reviewed journal, conference proceedings papers, and book chapters. She is a fellow of AIMBE and AAPM and the recipient of a 2014 R&D100 award.

Roger M. Engelmann is an application developer in the Department of Radiology at The University of Chicago. His work includes development of software to support CAD research as well as software to integrate research and clinical information systems.

Maryellen L. Giger is the A.N. Pritzker Professor of Radiology and the Committee on Medical Physics at The University of Chicago. Her research interests mainly involve the investigation of CAD and radiomics methods for the assessment of risk, diagnosis, prognosis, and response to therapy of breast cancer on multimodality (mammography, ultrasound, and magnetic resonance) images. She is also involved in broad-based developments in computer vision and data mining of medical images.

George Redmond is a program director for the NCI's Cancer Imaging Program. He played a key role in several successful large enterprise systems development initiatives at NCI. He implemented three enterprise systems to improve the clinical trial process to advance the restructuring of the nation's cancer clinical trials enterprise. He is the recipient of the prestigious NIH Director's Award for the successful development of an enterprise system for cancer therapeutic management and development.

Keyvan Farahani is a program director for the Image-Guided Interventions Branch of the NCI's Cancer Imaging Program. In this capacity, he is responsible for the development and management of NCI initiatives that address diagnosis and treatment of cancer and precancer through integration of advanced imaging and minimally invasive and noninvasive therapies. He has led the organization of brain tumor segmentation challenges at MICCAI 2013-2015. His graduate studies were in biomedical physics (UCLA, '93).

Justin S. Kirby is a bioinformatics analyst at the Frederick National Laboratory for Cancer Research. His focus is on developing informatics methods to improve reproducibility in imaging research through increased data and code sharing, as well as the adoption of structured reporting standards. His team manages The Cancer Imaging Archive, which provides free and open-access datasets of deidentified cancer images to researchers.

Laurence P. Clarke was a branch chief for the Imaging Technology Development Branch of the NCI's Cancer Imaging Program. He was responsible for the development of research strategies and initiatives that support emerging imaging technologies to address the cancer problem. His responsibilities included the development of web-accessible research resources for benchmarking the performance of imaging platforms and clinical decision tools. He oversaw several research networks for the development of quantitative imaging methods for multicenter clinical trials.