Journal of
Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

LUNGx Challenge for computerized lung nodule classification: reflections and lessons learned

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

SPIE.

LUNGx Challenge for computerized lung nodule classification: reflections and lessons learned

Samuel G. Armato III
University of Chicago
Department of Radiology
MC 2026
5841 S. Maryland Avenue
Chicago, Illinois 60637, United States
E-mail: s-armato@uchicago.edu

Lubomir Hadjiiski
University of Michigan
Department of Radiology
1500 E. Medical Center Drive
Ann Arbor, Michigan 48109, United States

Georgia D. Tourassi
Biomedical Science and Engineering Center
Health Data Sciences Institute
Oak Ridge National Laboratory
Oak Ridge, Tennessee 37831, United States

Karen Drukker
Maryellen L. Giger
Feng Li
University of Chicago
Department of Radiology
MC 2026
5841 S. Maryland Avenue
Chicago, Illinois 60637, United States

George Redmond
Keyvan Farahani
National Cancer Institute
Division of Cancer Treatment and Diagnosis
Cancer Imaging Program
9609 Medical Center Drive
Bethesda, Maryland 20892, United States

Justin S. Kirby
Frederick National Laboratory for Cancer Research
Leidos Biomedical Research, Inc.
Cancer Imaging Program
Frederick, Maryland 21702, United States

Laurence P. Clarke
National Cancer Institute
Division of Cancer Treatment and Diagnosis
Cancer Imaging Program
9609 Medical Center Drive
Bethesda, Maryland 20892, United States

Challenges, in the context of medical imaging, are valuable in that they allow for a direct comparison of different algorithms designed for a specific radiologic task, with all algorithms abiding by the same set of rules, operating on a common set of images, and being evaluated with a uniform performance assessment paradigm. The variability of system performance based on database composition and subtlety, definition of "truth," and scoring metric is well-known;[1–3] challenges serve to level the differences across these various dimensions. The medical imaging community has hosted a number of successful thoracic imaging challenges that have spanned a wide range of tasks,[4,5] including lung nodule detection,[6] lung nodule change, vessel segmentation,[7] and vessel tree extraction.[8] Each challenge presents its own unique set of circumstances and considerations; however, important common themes exist. Future challenge organizers (and participants) could benefit from an open discussion of successes achieved, pitfalls encountered, and lessons learned from each completed challenge.

The LUNGx Challenge, formally known as the SPIE-AAPM-NCI Lung Nodule Classification Challenge, was a collaborative effort sponsored and supported by the International Society for Optics and Photonics (SPIE), American Association of Physicists in Medicine (AAPM), and National Cancer Institute (NCI) along with investigators from University of Chicago, University of Michigan, and Oak Ridge National Laboratory. The Challenge was conducted as part of the SPIE Medical Imaging Symposium held in Orlando, Florida from February 22 to 26, 2015. The focus of the LUNGx Challenge was the computerized classification of lung nodules as benign or malignant in diagnostic computed tomography (CT) scans. This communication provides an overview of the LUNGx Challenge and addresses the "lessons learned" during the conceptualization, conduct, and analysis of the Challenge.

1 Approach

The Challenge included a calibration phase and a testing phase. A calibration set of 10 thoracic CT scans, five with a single confirmed benign nodule and five with a single confirmed malignant nodule, was made available through the NCI's The Cancer Imaging Archive (TCIA)[9] on November 26, 2014. Along with the complete set of DICOM images for these 10 clinical CT scans, a spreadsheet was included that contained the spatial coordinates of the approximate center of each nodule and the diagnosis (the "truth") for each nodule. All information within the DICOM headers remained intact with the exception of protected health information, which had been removed by the organizers prior to the upload of images to TCIA; this anonymization approach and the public release of the CT scans for this purpose had been approved by the local Institutional Review Board. The nodules in the calibration set (and the test set) had been determined by a radiologist to be either primary lung cancer or benign based on pathologic assessment and/or follow-up imaging examinations. As stated in the Challenge announcement,[10] the calibration set was meant to be representative of the technical aspects (e.g., scanner type, acquisition parameters, file format) associated with images in the test set so that participants could become familiar with the image acquisition parameters and DICOM file structure of the one institution (University of Chicago) that supplied the clinical cases for the LUNGx Challenge; participants were not to consider the lung nodules present in the calibration set to be representative of the difficulty level expected in the test set. It is important to note that the calibration set was not intended to serve as a development or training set, since the expectation was that participating groups already would have developed a trained system.

The test set became available through TCIA on January 12, 2015. The test set contained 60 thoracic CT scans with a total of 73 nodules (13 scans contained two nodules each). Along with the complete set of DICOM images for these 60 clinical CT scans, a spreadsheet was included that contained the spatial coordinates of the approximate center of each of these 73 nodules. All information within the DICOM headers remained intact with the exception of protected health information, which had been removed by the organizers prior to the upload of images to TCIA. Participants applied their algorithms to these 73 lung nodules to assign a probability of malignancy to each nodule.

Fifteen sets of results from participants' algorithms (in the form of a single numeric value estimate of the probability of malignancy for each nodule) were submitted to the organizers by February 6, 2015. With knowledge of the truth, the organizers evaluated the performance of each of these 15 sets of submitted results using receiver operating characteristic (ROC) analysis. Each group that submitted results was invited to prepare a poster for display at the SPIE Medical Imaging Symposium, and the two groups with the best performance (greatest area under the ROC curve) were invited to participate in a panel discussion at the Symposium entitled, "CAD grand challenge: present and future." In addition, one member of the highest-performing group was awarded complimentary registration to the Symposium.

2 Lessons Learned

2.1 *Establishing a Challenge*

The organizers of a challenge contribute time, effort, and resources, along with the ability to anticipate unforeseen situations. Organizers must establish the necessary controlled environment that includes a focused, well-vetted set of (clinical) cases on the front end and, on the back end, a performance assessment process that specifies the manner in which participants are to report results to the organizers and the scoring metric through which the results will be evaluated. Essential to the successful implementation of a challenge is a robust infrastructure for case dissemination, communication, and upload of consistently formatted results. The need for communication begins with the initial announcement; appropriate venues for "advertising" an upcoming challenge should be identified early in the planning process. The LUNGx Challenge made use of the SPIE Medical Imaging e-mail distribution list and the Grand Challenges in Biomedical Image Analysis web site.[11] During the Challenge, all questions from participants to the organizers were managed through Google Groups so that all participants received the same information and no one participant solely benefitted from clarifying information that might be conveyed in answer to a question.

Despite diligent planning and effort on the part of the organizers, a challenge will not succeed without the active and dedicated participation of groups willing to "accept the challenge." The phrase "build it and they will come" certainly applies in the setting of challenges; indeed, the burden on organizers is to "build it SO they will come" by offering an activity that participants consider reputable and one they consider worthy of their time, effort, and involvement; also attractive is some type of incentive (e.g., participation on a conference panel or co-authorship on an eventual publication). In the case of the LUNGx Challenge, sponsorship by three major organizations (SPIE, AAPM, and NCI) and affiliation with the SPIE Medical Imaging Symposium provided the legitimacy, while participation in the Symposium along with the potential for complimentary registration provided an incentive. Groups that choose to participate in a challenge deserve much credit for subjecting their algorithms to circumstances that might differ substantially from those under which the algorithms were developed.

2.2 *Database Considerations*

The single most important component of a medical imaging challenge is the set of images (the "database"); therefore, the effort involved in database collection should not be underestimated or undervalued. The organizers typically need to provide both a set of training cases (or a more limited set of calibration cases as was done for the LUNGx Challenge) and a set of test cases, which leads to the ubiquitous question, "How many cases do we need?" The answer to this question is complicated. The balance between the effort (cost) required to obtain relevant clinical images (which increases with increasing numbers of cases) and the statistical power that may be achieved by the challenge (which also increases with increasing numbers of cases) can be elusive, and often practicality emerges as the deciding factor, with the number of cases being determined by the (limited) volume of cases available to the organizers.

Collecting cases for a challenge requires consideration of a number of factors that impose constraints on the selected cases, thus increasing the burden of database collection—but with the expected benefit of a more scientifically or clinically relevant challenge. Two general aspects of the collected cases must be considered: (1) the distribution of image-acquisition parameters represented by the images and (2) the range of disease states or anatomic variation captured by the images. The LUNGx Challenge used clinical cases from one institution to minimize the inherent variability of scan technical parameters, although other challenges could benefit from the greater heterogeneity in imaging parameters and patient demographics captured by cases from multiple institutions. The organizers must determine the level of consistency across all image-acquisition parameters that is required by the challenge task; for some challenges, a specific image reconstruction algorithm, pixel size, section thickness, or contrast-enhancement protocol might be desired, while for other challenges, a clinically realistic distribution of these parameters could be more appropriate. For challenges that involve evaluation of abnormalities, the range of lesion size and subtlety is an important consideration, and the distributions of factors such as gender and age should not be overlooked, since, depending on the challenge task, these factors can have a substantial impact on the results. If the challenge task is one of discrimination between two conditions, gender and/or age matching between groups might be necessary. The nodules between the two groups in the LUNGx Challenge test set were size matched (although this fact was not disclosed to participants), since nodule size is a well-known predictor of malignancy; had size matching not been implemented, participants potentially could have achieved a high level of performance simply by calculating some metric of nodule size alone.[12] Organizers should verify that data available in the DICOM image headers or in any supplemental documentation does not incidentally yield a high performance in the specified task. Other clinical factors might be relevant to the challenge task, such as smoking history, race, or genetic information, and the organizers must decide whether such additional clinical or demographic data is essential to the challenge. For the LUNGx Challenge, some participants may have reasonably desired smoking history, the distribution of cell type among the malignant nodules, or the processes represented by the benign nodules, but this information was not provided.

2.3 *Clarity of Challenge Rules*

Based on their own preconceptions and extensive experiences in the field, challenge organizers likely will have certain expectations with regard to the manner in which participants should approach the challenge task; these expectations, without a doubt, will be unwittingly violated by some participants who come from different technical backgrounds/cultures or who have a different interpretation of the challenge rules or even the fundamental task that the challenge seeks to address. In the LUNGx Challenge, for example, the 10 calibration cases were intended to assist groups evaluate the compatibility of the Challenge cases with their algorithms and were not intended for algorithm development or for classifier training; nevertheless, some groups attempted to use this intentionally small set of calibration cases for development or training. As another example, despite the expectation that the nodule classification systems would be fully computerized with no human involvement, some groups sought out and incorporated local radiologist input: radiologist-constructed nodule outlines, radiologist semantic characterization of nodules, and radiologist ratings of nodule malignancy were used as input to the systems that some participants applied to the Challenge test set (the first two uses of human involvement were ruled to be acceptable—although unexpected—systems, while the system that used radiologist malignancy ratings was withdrawn). The rules of a challenge must be crafted as completely, as clearly, and as logically as possible, with the organizers attempting to anticipate any confusion and misinterpretations the rules might cause. After numerous inquiries from groups working with the LUNGx calibration set, the Challenge rules were expanded to explicitly define the spatial coordinate system conventionally used for CT scan images and to describe basic elements of the DICOM file format.

Any group that downloaded cases while the challenge was active should be allowed to retain those images; the alternative (i.e., requesting that groups delete all downloaded data at the conclusion of the challenge) is impractical and unenforceable. Accordingly, plans for the continued public availability of challenge cases should be made, since cases from a challenge can provide a valuable resource. With this in mind, the LUNGx Challenge instructions stated that "anyone wishing to use the downloaded images for presentation or publication purposes outside of the LUNGx Challenge should acknowledge the SPIE, the NCI, the AAPM, and the University of Chicago. The LUNGx Challenge cases and associated data may be downloaded from Ref. 13. One lingering issue with the LUNGx Challenge going forward is whether the truth information for the test cases will be made public; knowledge of the nodule diagnosis of each test case would greatly enhance the value of these cases as a resource for medical imaging researchers, but disclosure of this information would exclude the potential incorporation of these test cases in a future challenge. Digital object identifiers (DOIs) have received growing interest as a more permanent mechanism through which to foster reproducible imaging research;[14] a recent lung segmentation challenge conducted by the

Quantitative Imaging Network made use of the DOI approach to maintain challenge data for future use.[15]

Participants should be made aware of their potential involvement in any presentation or publication expected to result from a challenge. The ability of participants to remain anonymous in a publication should be stated in the challenge rules. Anonymity could take one of two general forms: (1) complete omission of the names and affiliations of participating groups or (2) a listing of participants that remains disassociated from the results of individual systems. A publication preferably should report the general methodology of each system linked with the performance results of that system to convey the relative merits of different approaches to the challenge task; under this approach, however, it may be impractical to maintain separation of participants' identities from their methods, and hence their results. Ultimately, full disclosure, full credit, and full responsibility are always preferred in a scientific communication—an argument that favors no level of anonymity at all.

2.4 *Participant Responsibility*

Groups that choose to participate in a challenge have a responsibility to approach the challenge with commitment and scientific rigor. Participants must adhere to the rules of the challenge, which should be clearly specified by the organizers. The point of commitment also should be specified. Download of the cases should not commit a group to participation in the challenge; once downloaded and assessed, the challenge images may be determined by a group to comprise, for example, technical parameters or concomitant disease for which the group's system was not designed. With submission of its system's output to the challenge organizers for evaluation, however, a group becomes a fully vested challenge participant and should accept the final results and performance analysis of the organizers. The spirit of a challenge is compromised (and resources of the organizers are wasted) if groups are allowed to withdraw their participation if they find their system's performance is not to their satisfaction. Participants must be mindful of the educational, friendly competition, and community-building nature of a challenge.

To summarize, the LUNGx Challenge was a successful scientific challenge for the computerized classification of lung nodules on CT scans. Despite careful planning and an attempt to think through potential pitfalls, some aspects of the Challenge yielded unexpected outcomes, but these situations provide valuable lessons for members of the medical imaging community who would organize (and participate in) future challenges. A scientific paper analyzing the individual and collective performances of algorithms submitted to the Challenge, along with ancillary analyses on clinically relevant subsets of the LUNGx database and radiologists' classification performance, is being prepared. The Challenge was given a prominent role in the CAD Conference of the 2015 SPIE Medical Imaging Symposium. Conference organizers plan to leverage the success of the LUNGx Challenge into a related challenge to be conducted in association with SPIE Medical Imaging 2016.

Acknowledgments

The authors would like to express their appreciation to Angela Keyser 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 the SPIE, NCI, and AAPM. 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 U.S. Government. S.G.A. and M.L.G. receive royalties and licensing fees through The University of Chicago related to computer-aided diagnosis. 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 United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a nonexclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States 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.[16]

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 et al., "The effect of verification on the assessment of imaging techniques," *Invest. Rad.* **18**, 194–198 (1983).
4. http://grand-challenge.org/Why_Challenges/
5. K. Murphy, "Development and evaluation of automated image analysis techniques in thoracic CT," PhD Thesis, Utrecht University, the Netherlands (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**, 707–22 (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**, 1217–32 (2014).
8. P. Lo et al., "Extraction of airways from CT (EXACT'09)," *IEEE Trans. Med. Imaging* **31**, 2093–2107 (2012).
9. K. Clark et al., "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository," *J. Digital Imaging* **26**, 1045–1057 (2013).
10. http://spie.org/Documents/ConferencesExhibitions/MI/LUNGxChallengeFormat.pdf.
11. http://grand-challenge.org/.
12. A. C. Jirapatnakul et al., "Characterization of pulmonary nodules: effects of size and feature type on reported performance," *Proc. SPIE* **6915**, 69151E (2008).
13. S. G. Armato, III et al., "SPIE-AAPM-NCI lung nodule classification challenge dataset," *Cancer Imaging Arch.* (2015).
14. P. E. Bourne, "DOIs for DICOM raw images: enabling science reproducibility," *Radiology* **275**, 3–4 (2015).

15. J. Kalpathy-Cramer et al., “QIN multi-site collection of Lung CT data with nodule segmentations,” *Cancer Imaging Arch.* (2015).
16. http://energy.gov/downloads/doe-public-access-plan.

Samuel G. Armato III is an associate professor in the Department 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 radiotherapy-induced normal tissue complications, and the automated detection of pathologic change in temporal subtraction images.

Lubomir Hadjiiski is a research 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 computer-aided diagnosis, 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 received her 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.

Karen Drukker has been actively involved in computer-aided diagnosis/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.

Maryellen L. Giger is the A. N. Pritzker Professor of Radiology and is on the Committee on Medical Physics at the University of Chicago. Her research interests mainly involve the investigation of CAD and radiomic 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.

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 on chest CT scans or chest radiography, computer-aided diagnosis, advanced imaging techniques, and tumor response assessment.

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 (http://www.cancerimagingarchive.net/), which provides free and open-access data sets of deidentified cancer images to researchers.

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