Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Special Section Guest Editorial: Evaluation Methodologies for Clinical AI

Susan M. Astley
Weijie Chen
Kyle J. Myers
Robert M. Nishikawa

SPIE.

Susan M. Astley, Weijie Chen, Kyle J. Myers, Robert M. Nishikawa, "Special Section Guest Editorial: Evaluation Methodologies for Clinical AI," *J. Med. Imag.* **7**(1), 012701 (2020), doi: 10.1117/1.JMI.7.1.012701

Special Section Guest Editorial: Evaluation Methodologies for Clinical AI

Susan M. Astley,[a] Weijie Chen,[b] Kyle J. Myers,[b,*] and Robert M. Nishikawa[c]

[a]University of Manchester, Division of Informatics, Imaging & Data Sciences, Manchester, United Kingdom

[b]U.S. Food and Drug Administration, Office of Science and Engineering Laboratories, CDRH, Silver Spring, Maryland, United States

[c]University of Pittsburgh, Department of Radiology, Pittsburgh, Pennsylvania, United States

With the explosion of deep learning applications in medical imaging there is an urgent need to develop methods to evaluate the performance of artificial intelligence (AI) systems due to the increased complexities/varieties of AI technologies, the dependence of these new technologies on large datasets, and the emergence of novel types of clinical applications of AI systems. Proper testing methodology, metrics, appropriate training/tuning/validation study designs, and statistical analysis methods are needed to ensure that studies produce meaningful, robust, and generalizable results in a least burdensome fashion. These elements are key to the clinical adoption of AI technologies. Thus this Special Section for the *Journal of Medical Imaging*, Volume 7, Issue 1, encouraged relevant submissions in these topic areas.

AI is not new to medical imaging. Since the earliest days of the SPIE Medical Imaging symposium there have been presentations on what was then referred to as Computer-Aided Diagnosis (CAD). The Computer-Aided Diagnosis conference at the larger SPIE Medical Imaging (MI) symposium was launched in 2006. Applications for CAD in mammography, lung CT, and chest x-ray imaging, all mature commercial products today, were discussed in their earliest phases at this conference. SPIE MI has also been the home for the introduction of new approaches to the assessment of CAD algorithms, a tradition that continues primarily through the conference on Image Perception, Observer Performance, and Technology Assessment. A perusal of the SPIE MI program through the years allows the reader to see the progression of AI algorithm development as well as methods for AI assessment.

What *is* new to AI is the recent advance in computational power and the availability of large datasets that have enabled the successful application of deep neural network (DNN) architectures for various medical imaging tasks. These tasks include the common applications in the field related to the finding of suspicious areas in images for a reader to give a second-look, as well as the characterization of a reader's identified areas of suspicion with the support of AI. Newer tasks to which DNNs are being applied include image denoising, full image reconstruction from highly sparse or very noisy projections, triage systems that alert the user to high-priority cases so as to adjust case-reading order, AI-selected image acquisition parameters on a per-patient basis, and the approximation of the ideal observer for use as a measure of image quality in complex imaging scenarios. For some applications, the performance of AI is being demonstrated to reach or surpass expert human performance such that automated diagnosis in which the clinician is replaced by the AI system is arguably close at hand. Moreover, the range of imaging modalities for which AI is being applied is vast, from the x-ray applications listed above, to optical, ultrasound, MRI, and digital pathology, the latter recently introduced as a conference track of its own at the SPIE MI symposium.

Across the wide and diverse landscape of AI applications and indications, there is a need for AI algorithm evaluation methods that accurately estimate the device performance generalizable to the clinic. Methods are needed to assess AI systems intended for use beyond the standard paradigm of AI as an aid or second reader. We need methods for determining that an AI system might be reliably used to rule out images from physician's review (that is, to partially replace clinicians), as well as for fully automated diagnosis (without human involvement). The community needs to develop

*Address all correspondence to Kyle J. Myers, E-mail: Kyle.Myers@fda.hhs.gov

consensus on appropriate methods for evaluating algorithms that detect a large number of different types of abnormalities from a single image or case; we know of some AI systems that are intended to find on the order of 10s to close to 100 different findings. How would a study be designed to include a sufficient number of patients for each type of finding and what would be the statistical method of analysis? In the area of CAD triage systems, how do we evaluate the accuracy and value of a CAD triage system and how do we take into account the risk that might be associated with the deployment of multiple such systems in an institution for different applications (cardiac plus stroke, for example) which might lead to triage-system conflict or alarm fatigue? Another novel type of application is AI-guided image acquisition, for instance, to have AI algorithms select imaging protocols for personalized image acquisition (e.g., MRI) or provide real-time guidance to help less experienced operators acquire images of sufficient diagnostic quality in the right views (e.g., echocardiography). These are new evaluation questions that are in the uncharted territory, have some methods in use but with little consensus, or need more research to make the evaluation more efficient.

The problem of determining the reference standard for cases/images for training and validation of AI is an old problem, but there are new challenges being brought to the fore as the number of types of findings from AI algorithms increases and AI systems are designed to estimate more complex entities, such as disease risk or case-reading priority. In the example of systems that identify 10s of findings, how is 'truth' determined for each finding on each case, or lack thereof? Some imaging applications, digital pathology being an excellent example, serve as the reference standard for upstream imaging and AI systems; AI applied to applications like digital pathology faces the challenge of determining a reference standard because there is no gold standard unless perhaps we are willing to wait for patient outcomes. Methods have been suggested for combining interpretations from multiple expert readers in such applications where a more objective reference standard is too costly or otherwise unobtainable, but consensus is lacking on the appropriate approach to take for a given AI application.

Once the performance of the AI algorithm itself has been estimated and determined to have clinical potential, the effect of the AI algorithm clinically needs to be assessed. There are standard methods for doing this in controlled situations, such as observer studies (simulated clinical reading environment), and in controlled clinical practice, such as randomized controlled trials or case-control studies, but ultimately, we need to know when implemented widely, whether patient care has been improved. It has been shown[1] that in wide clinical use, systems designed to assist radiologists read screening mammograms are neither used as they were designed nor how they were used in testing. As a result, the effectiveness of the technique was compromised.

Somewhat ironically, while AI developers are often trying to make radiologists more accurate, radiologists may use AI tools for reading efficiency. This may not be surprising because when reading in a stressful, heavy workload environment, where the number of cases read is the most immediate feedback, time management is critical. Thus while an AI system may improve radiologists' performance in clinical trials, where there is not the same stress and time constraints, they may fail to do so in actual day-to-day clinical use.

Increases in computing power have also enabled the increased availability of accurate simulations of patients and imaging systems. Such simulations allow for the use of "virtual patients" with known truth for data augmentation in AI training. Such simulations can support the evaluation of the robustness of AI algorithms to different image acquisition protocols and image quality attributes as well. Data simulations also provide a key resource in the development and evaluation of new statistical methods for the assessment of AI, for example, the comparison of the efficiency of competing study designs. We are strong advocates for research investment into the development and sharing of computational modeling and simulation tools for these purposes.

Continuously learning AI systems have become commercially available in other industries, although not widely embraced and available for use in clinical imaging applications. Our field needs methods and metrics for assessing algorithm adaptation protocols that would give confidence in a continuously learning system's ability to improve with additional access to cases such that the burden of repeated validation steps might be reduced.

The goal of this JMI special section is to publish recent research on evaluation methodologies and clinical studies that will help to define proper methods and consensus on best practices for evaluating AI systems clinically. This is a high bar, when one considers the stages in the evaluation of an AI algorithm. In the first stage, the algorithm will be tested in a stand-alone manner

Table 1 Summary of practical-application papers in the special section.

Author, paper title	Imaging modality and clinical task	Reference standard	Sample size for training/ tuning	Sample size for validation	Performance metrics
Cha et al., Evaluation of data augmentation via synthetic images for improved breast mass detection on mammograms using deep learning	X-ray mammography; breast mass detection	Radiologist outlining/simulation	1231 mammograms (1318 masses) and 2000 synthetic images	361 mammograms (378 masses)	Area under the ROC curve (AUC)
Saadeh et al., Histopathologist-level quantification of Ki-67 immunoexpression in gastroenteropancreatic neuroendocrine tumors using semiautomated method	Digital pathology; Ki-67 biomarker quantification	Manual counting by three pathologists	n/a (publicly available image quantification tool ImageJ)	20 cases	Intra-class correlation coefficient; concordance correlation coefficients; Bland-Altman plot
Gudmundsson et al., Deep learning-based segmentation of malignant pleural mesothelioma tumor on computed tomography scans: application to scans demonstrating pleural effusion	CT; tumor volume segmentation	Manual segmentation by one radiologist	2663 CT sections from 76 scans of 61 patients	Set 1: 94 CT sections from 46 scans of 34 patients Set 2: 130 CT sections from 43 scans of 43 patients	Dice similarity coefficient
Schau et al., Predicting primary site of secondary liver cancer with a neural estimator of metastatic origin	Digital pathology; predicting primary site of secondary liver cancer	Clinical annotation	180 slides	51 slides	F1 score and other diagnostic metrics
Whitney et al., Harmonization of radiomics of breast lesions across international DCE-MRI datasets	Dynamic contrast-enhanced (DCE-MR); distinguishing between cancers and benign lesions	Molecular subtype from pathology	US: 680 lesions China: 1549 lesions	10-fold cross-validation	Area under the ROC curve (AUC)

(how the algorithm performs on its own, and not in terms of its impact on reader performance) on limited cases in the hands of the developer. When the algorithm is submitted for regulatory review by the FDA, a much larger dataset is typically required to demonstrate that the algorithm generalizes to multiple clinical sites and across the heterogeneity of the wider patient population contained in the indications for use. Evidence that the benefit of the device outweighs the risks typically involves demonstrating that the algorithm improves reader performance, though the study to show this is often a "lab study" in which cases may be enriched; the reader may not have common metadata including patient age, symptoms, or prior scans; and the case interpretation without and with the AI is done retrospectively, that is, without impact on patient management. In other words, even at the phase of FDA submission there can be unanswered questions regarding the performance of the AI algorithm when used in routine clinical practice. The true performance of an AI system in the clinic will become known over time as the algorithm is applied to the actual patient population. We all seek better (more efficient and more accurate) approaches to the evaluation of AI that predict that ultimate clinical performance, so that algorithms implemented in the clinic can be counted on to perform as expected, improving the lives of patients as intended. The papers in this special section are contributions toward this goal.

This special section contains six articles that include both a theoretical perspective and practical applications for the evaluation of AI systems in medical imaging. Barrett reviews several basic concepts from image science that, from the author's perspective, may be useful in designing and validating AI-based imaging systems.[2] The state-of-the-art deep-learning-based AI is commonly viewed as a "black box," which has been empirically found to provide useful solutions to many practical problems, but the scientific mechanisms behind these solutions are not well understood. The theoretical perspectives provided in Barrett's article encourage further research in understanding AI in connection with image science wisdom. Understanding AI is not merely a dream in the ivory tower but may have important practical implications, as explainable AI may be more easily translated to the clinic than a "black box."

The other five articles in this special section, as summarized in Table 1, are practical applications in both radiology and digital pathology and shed light on many aspects of evaluation methodologies. The article by Cha et al. shows the potential usefulness of data augmentation with synthetic images in training data-hungry AI algorithms; it also demonstrates the importance of realism of augmented data for such data to be useful. The articles by Saadeh et al., Gudmundsson et al., and Schau et al. apply AI to different clinical tasks (segmentation, quantification, classification) and all rely on a reference standard established by human observers. It is evidently important to develop methods for accounting for uncertainties from such reference standards in the evaluation of AI performance. Furthermore, these three studies are at different phases of research, ranging from a feasibility study to incremental improvement and a dedicated assessment of a public image analysis tool. Together they show that proper evaluation methods are not only critical in the final stage of translating a technology to the clinic, but useful during development/refinement of a technology. Finally, the paper by Whitney et al. demonstrates a method for the harmonization of radiomic features when databases from different institutions are combined, potentially enabling computer-aided diagnosis models that are robust to variations in image acquisition and processing differences across imaging sites.

We appreciate and congratulate all the authors contributing to this special section. We hope these papers and this editorial help the readers acquire a sense of the scope, challenges, and opportunities in the evaluation of AI technologies for medical imaging applications. We look forward to future articles on these topics, thereby developing consensus on appropriate methodologies to translate safe and effective AI technologies to the clinic to benefit patients.

References

1. R. M. Nishikawa and K. T. Bae, "Importance of better human-computer interaction in the era of deep learning: mammography computer-aided diagnosis as a use case," *J. Am. Coll. Radiol.* **15**(1), 49–52 (2018).
2. H. H. Barrett, "Is there a role for image science in the brave new world of artificial intelligence?" *J. Med. Imaging* **7**(1), 012702 (2019).

Biographies of the authors are not available.