OPINION

Best Practices for Artificial Intelligence and Machine Learning for Computer-Aided Diagnosis in Medical Imaging

Daniel Vergara, MS, Samuel G. Armato III, PhD, Lubomir Hadjiiski, PhD, Karen Drukker, PhD, CADSC (TG 273)

INTRODUCTION

The American Association of Physicists in Medicine Task Group (TG) 273 has been charged with developing recommendations on best practices for the development and performance assessment of computer-aided decision support systems. The TG report [1] addresses broad issues common to the development of most, if not all, computer-aided diagnosis (CAD) and artificial intelligence (AI) applications and their translation from the bench to the clinic. The goal was to bring attention to issues such as proper data collection and training and validation methods for machine learning (ML) algorithms, aiming to improve generalizability and reliability and thus accelerate the adoption of CAD-AI systems for clinical decision support. The report focuses on several developmental stages of CAD-AI: data collection, reference standards, model development, performance assessment, and translation to the clinic.

CAD is the use of computer-analyzed information to assist in medical decision making. Traditional ML CAD was introduced into radiology clinical practice more than two decades ago. In recent years, rapid advances in ML and deep learning techniques have given rise to the development of CAD tools incorporating AI and dramatically increased interest in these methods for clinical decision support.

DATA COLLECTION AND ETHICS

Proper data collection methods are critically important to the successful training, validation, and implementation of CAD-AI algorithms. Improper collection and manipulation of data (eg, improper data augmentation) can lead to an overestimation of performance or lack of generalizability.

Data collected should reflect the intended use and population to allow for the replication of results in a real-world clinical setting. Population demographics, ethics, case sampling strategies, and sample sizes must be carefully considered. Training data could be collected with methods such as stratified sampling for improved case balance (eg, cases with or without disease or different racial/ethnic groups). Improper data collection practices may introduce bias and create misleading model performance, especially in subpopulations that may not be appropriately represented in the study data set. It is also of critical importance to use multi-institutional data with diverse patient demographics and equipment and protocol variability for training to improve model generalizability. Public data sets from multiple sites help alleviate bias due to limited data diversity at any individual site. To create public data sets, care must be taken to ensure transparency on inclusion and exclusion criteria, image acquisition parameters, and proper deidentification of protected health information. Image quality and reference standard integrity should be verified before a data set is released. Effective methods to preserve sequestered data for testing with unseen cases are important for reliable evaluation and fair comparison of CAD-AI algorithms.

Data collected from different clinical sites can vary, potentially leading to undesirable systematic variations in the CAD-AI output. Harmonization can be used to reduce these variations retrospectively after acquisition while preserving the biological variability captured in the images. Although harmonization methods usually cannot address the issue of systematic variations among patient subpopulations, they aim to reduce the systematic variations due to differences in image acquisition equipment, protocols, reconstruction, and postprocessing among data collection sites. Harmonization methods include image- and feature-domain harmonization. Image-domain harmonization includes postprocessing of image data, and feature-domain

1546-1440/24/$36.00 ■ https://doi.org/10.1016/j.jacr.2023.10.021

harmonization includes statistical normalization techniques.

Rapid advances in image acquisition hardware and software can lead to data set obsolescence. To create an enduring image data set, data collection and management should be considered a continuous process to ensure that the images were acquired with equipment that is still technically relevant and in accordance with appropriate image acquisition protocols.

Data augmentation is a collection of task-dependent techniques used to improve performance and generalization of CAD-AI by increasing the training set size, usually by adding altered training data (eg, rotation or scaling) or by introducing synthetic or generative data. Any such changes should not modify the appearance of the image in a manner that distorts underlying anatomy and biology beyond clinical reality. Data augmentation is not equivalent to increasing the number of independent cases in the data set, and proper validation of such trained models is crucial.

REFERENCE STANDARDS

A reference standard [1] is required to train CAD-AI in a supervised learning setting and to evaluate its performance. The utility of a CAD-AI algorithm is critically linked to the quality of the reference standard used to develop it. Methods for acquiring a reference standard (annotations) include expert labels, electronic health record, crowd sourcing, and weak or noisy labels [1]. An objective reference standard supported by complete and reliable clinical and pathological data is preferred. When a subjective reference standard cannot be avoided (eg, lesion segmentation), independent assessments by multiple domain experts should be obtained, and their variabilities should be evaluated. Potential biases that may be introduced when training or evaluating an algorithm with such reference standards should be assessed. Given this potential variability in the reference standard, the use of the terms "ground truth" or "gold standard" is discouraged [1-5].

MODEL DEVELOPMENT

When developing a CAD-AI model, it is important to consider factors that will affect the robustness of a model and minimize the risk of overfitting to the training set, such as data sampling, the levels of learning supervision, and training strategies. Data sets should generally be split into three nonoverlapping partitions: training, validation, and testing. It is of critical importance to use an independent test set representative of the intended use that has not been employed in model training or model optimization for final performance evaluation. This test set should be used only *once* to report the final performance level; ideally, the test set should be sequestered from model developers because multiple uses of the test set by the same developer will introduce bias and thus weaken generalizability.

The ML strategy (eg, transfer learning, federated learning, and continuous learning) must also be considered before training a CAD-AI model. These ML strategies provide various levels of supervision: supervised, semisupervised, self-supervised, unsupervised, and multiple instances. Transfer learning can be implemented by training a network on a source task and then using the pretrained weights to initialize the training for a target task, rather than random initialization. Transfer learning includes multitask learning and domain adaptation. Federated learning enables collaborative training on decentralized data sets. A continuous learning system adapts over time to the changing environment; although appealing, effective methods should be implemented to monitor the performance of such systems in clinical settings to safeguard its reliability.

PERFORMANCE ASSESSMENT

Selection of performance metric(s) depends on the task and the reference standard. Often multiple performance metrics are appropriate and frequently desirable. Common performance metrics include receiver operating characteristic analysis, free-response receiver operating characteristic analysis, sensitivity, specificity, and Dice coefficient for segmentation. Performance analysis should include error estimates, assessment of statistical significance, and preferably assessment of reproducibility (eg, technical, statistical, inferential).

The intended use of a CAD-AI system must match the clinical environment in which it will be deployed. Multireader multicase studies with a data set representative of the intended patient cohort can be used to estimate the clinical impact of a CAD-AI algorithm.

TRANSLATION TO THE CLINIC

Translation of a CAD-AI system to the clinic requires approval by the appropriate regulatory body, an efficient user interface, acceptance testing, adequate user training, and robust postdeployment quality assessment (QA) procedures to monitor the consistency of performance over time.

The human-machine interface can impact the usefulness and acceptance of a CAD-AI tool for clinical use. The "black box" nature of AI software makes it difficult to understand the capabilities and limitations of an algorithm. Explainable AI is a rapidly evolving ML topic that seeks to increase the confidence of physicians using CAD-AI tools. The explanation must be consistent with medical knowledge or supported by clinical evidence. The most common approaches at present include generating visual heat maps, providing examples of similar lesions or cases, and providing semantic textual

explanations or cues. However, these approaches often cannot meet key requirements for utility and robustness. Additional development and validation are need before clinical use. For clinical tasks more complicated than lesion detection, the CAD-AI tool may need to provide explanations or references that correlate the recommendation with the patient's medical conditions or other clinical data [1].

Acceptance testing must precede clinical use of any CAD-AI tool. Manufacturers must provide detailed guidance on system installation, acceptance testing, and periodic QA. They should also provide specifications of performance levels and tolerance limits.

Proper use of a CAD-AI tool in the clinical workflow must be clearly understood. An initial user training phase followed by an adjustment phase is recommended as an integral part of CAD-AI deployment. During the adjustment phase, physicians should evaluate the performance of the tool in their patient population without being influenced in their clinical decisions to allow the physician to gain an appropriate level of confidence in the CAD-AI tool. More advanced validation involves prospective clinical assessments of the impact of CAD-AI on clinical outcomes using well-designed clinical trial protocols.

Prospective surveillance and periodic QA are recommended after initial clinical implementation. Use of phantoms for QA may be possible for some specific applications, although, in general, QA for CAD-AI will require human data. A fixed data set may be used to monitor changes in the CAD-AI system, but shifts in the imaging characteristics will necessitate continuously updated clinical data. Practical and effective QA procedures and metrics should be established as an integral part of clinical translation of CAD-Al.

CONCLUSION

The recommendations in the TG-273 report cover essential elements of CAD-AI systems. Considering the best practices during system development, validation, and deployment can help improve the generalizability and successful clinical translation of such systems.

REFERENCES

1. Hadjiiski L, Cha K, Chan H-P, et al. Recommendations on best practices for AI and machine learning for computer-aided diagnosis in medical imaging. Med Phys 2023;50:e1-24.
2. Armato SG 3rd, Huisman H, Drukker K, et al. PROSTATEx challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. J Med Imaging (Bellingham) 2018;5:044501.
3. Gold R, Reichman M, Greenberg E, et al. Developing a new reference standard: is validation necessary? Acad Radiol 2010;17:1079-82.
4. Genders TS, Ferket BS, Hunink MG. The quantitative science of evaluating imaging evidence. JACC Cardiovasc Imaging 2017;10:264-75.
5. Petrick N, Sahiner B, Armato SG III, Bert A, et al. Evaluation of computer-aided detection and diagnosis systems. Med Phys 2013;40 087001-1,17.

Daniel Vergara, MS, is from the Department of Radiology, University of Washington, Seattle, Washington; Member of American Association of Physicists in Medicine Computer-Aided Image Analysis Subcommittee and Task Group 273. Samuel G. Armato III, PhD, is from the Department of Radiology, The University of Chicago, Chicago, Illinois; Committee on Medical Physics Chair; Graduate Program in Medical Physics Director; Human Imaging Research Office Faculty Director; Associate Director for Education, University of Chicago Comprehensive Cancer Center; Treasurer, American Association of Physicists in Medicine; Vice Chair, American Association of Physicists in Medicine Computer-Aided Image Analysis Subcommittee and Task Group 273; Member of American Association of Physicists in Medicine Medical Imaging and Data Resource Center Subcommittee; and Member of International Society for Optics and Photonics (SPIE). Lubomir Hadjiiski, PhD, is from the Department of Radiology, University of Michigan, Ann Arbor, Michigan; Chair of American Association of Physicists in Medicine Computer-Aided Image Analysis Subcommittee and Task Group 273; Chair of National Institutes of Health National Cancer Institute Quantitative Imaging Network; Member of American Association of Physicists in Medicine Medical Imaging and Data Resource Center Subcommittee; Vice Chair of American Association of Physicists in Medicine Working Group on Grand Challenges; Member of American Association of Physicists in Medicine Imaging Physics Committee; Member of International Society for Optics and Photonics (SPIE), and Member of Institute of Electrical and Electronics Engineers (IEEE). Karen Drukker, PhD, is from the Department of Radiology, The University of Chicago, Chicago, Illinois; Member of American Association of Physicists in Medicine Computer-Aided Image Analysis Subcommittee and Task Group 273; Member of American Association of Physicists in Medicine Medical Imaging and Data Resource Center Subcommittee; Chair of American Association of Physicists in Medicine Working Group on Grand Challenges; Member of American Association of Physicists in Medicine Science Council; and Member of International Society for Optics and Photonics (SPIE). CADSC (TG 273): American Association of Physicists in Medicine Computer Aided Image Analysis Subcommittee (Task Group 273) with members: Kenny Cha, Heang-Ping Chan, Karen Drukker, Lia Morra, Janne J. Näppi, Berkman Sahiner, Hiroyuki Yoshida, Quan Chen, Thomas M. Deserno, Hayit Greenspan, Henkjan Huisman, Zhimin Huo, Richard Mazurchuk, Nicholas Petrick, Daniele Regge, Ravi Samala, Ronald M. Summers, Kenji Suzuki, Georgia Tourassi, Amita Shukla-Dave, Usman Mahmood, Daniel Vergara, Samuel G. Armato III (Vice Chair), Lubomir Hadjiiski (Chair).

Disclosure: This work represents the opinion of the authors and not necessarily that of Department of Health and Human Services or National Institutes of Health.

Karen Drukker receives royalties from Hologic Inc. The other authors state that they have no conflict of interest related to the material discussed in this article. The authors are non-partner/non-partnership track/employees.

Lubomir Hadjiiski, PhD: University of Michigan, Department of Radiology, MIB C476, 1500 E Medical Center Dr, Ann Arbor MI 48109; e-mail: lhadjisk@umich.edu