Failures Hiding in Success for Artificial Intelligence in Radiology

Saptarshi Purkayastha, PhD, Hari Trivedi, MD, Judy Wawira Gichoya, MD, MS

INTRODUCTION

Reports of computer algorithms outperforming radiologists have persisted over the last 15 years, starting with the 2005 publication by Rubin et al on detecting pulmonary nodules from CT scans [1]. Back then, these technologies were referred to as computer-aided diagnosis, which could be considered as a precursor, of sorts, to what is now referred to broadly as artificial intelligence (AI). Technology gains in hardware over the past 5 years have facilitated the training of deep neural networks with millions of parameters, exponentially accelerating the pace of AI publications. However, like every other scientific field, successes of AI in radiology are published and publicized with much fanfare, and failures are not discussed or made public. In fact, most AI failures are discovered anecdotally from personal experience or when shared in social media as tweets or blog posts. In this article, we discuss some pitfalls frequently encountered in reporting the success of AI in radiology, which might be considered failures when considered differently.

COMPARISON WITH HUMAN PERFORMANCE

A review of publications that describe the success of AI in diagnostic radiology highlights that they often exclude well-known knowledge that impacts human performance. Consider the utilization of AI to improve the diagnostic performance of screening mammograms. The 2013 article by Rafferty et al at five large academic health centers showed that multireader, multimodality (mammography and tomosynthesis) algorithms significantly improved human diagnostic accuracy [2] when interpreting screening mammograms. These findings have been increasingly adopted into practice, including insurance coverage of tomosynthesis for breast cancer screening in the United States and use of multiple readers for interpretation in the United Kingdom. However, a large-scale, multicountry study comparing the performance of an AI system with 101 radiologists failed to evaluate model performance against the same double-reader performance and multimodality (mammography and tomosynthesis) standard, despite knowledge that this would improve human performance [3].

If a radiology AI algorithm is to be evaluated at its optimal performance parameters, then so should the radiologists against which it is compared. Moreover, such retrospective reader studies often consist of enriched test sets, resulting in a well-understood laboratory effect that causes underperformance by radiologists because of a significant deviation from what a radiologist would encounter clinically. Many studies also ensemble multiple AI algorithms to improve performance, but human performance is not measured as double-read studies.

These types of studies have become the norm in human versus machine in radiology and are touted as successes; instead, they should be considered as potential misrepresentations and reported as such. Alternatively, recent studies have incorporated factors that influence human performance to compare human plus machine with AI alone. Consider the widely publicized publication by McKinney et al. These authors found that the AI algorithm outperformed all six individual radiologists. However, they then further simulated a double-reading partnership between the AI and radiologists, similar to current practice in the UK, and demonstrated equivalent performance to double-radiologist reads and reduced reading time by 88% [4]. We hope more articles consider human factors when comparing AI with radiologists to provide a more accurate and common playing field for AI in clinical practice.

HIDDEN STRATIFICATION OR INCOMPLETE SET OF LABELS

Another common area of failure of AI models is when reportedly top-

1546-1440/20/$36.00 ■ https://doi.org/10.1016/j.jacr.2020.11.008

performing models actually underperform on data sets that are labeled incompletely or differently from the data on which the model was trained. These are problems of generalizability, caused by hidden stratification. Hidden stratification occurs because AI performance measures are dominated by larger subsets; hence, if there are unrecognized subsets in the data, then the model performance drops [5]. For example, the musculoskeletal x-ray MURA (musculoskeletal radiographs) data set had binary labels of "normal" and "abnormal," and deep learning models trained on this data set had an aggregate receiver operating characteristic (ROC) area under the curve (AUC) of 0.91. However, when the data were relabeled with three subclass identifiers, the ROC AUC for the degenerative joint disease and fractures fell to 0.76 and 0.86, respectively, whereas AUC for hardware was higher at 0.98 [5]. Data sets that generate labels using natural language processing algorithms (eg, the National Institutes of Health Chest Xray14) are also commonly prone to stratification problems, and AI models that are trained on these data sets further propagate this intrinsic misrepresentation of the data. For example, 86% of cases reviewed in the National Institutes of Health Chest Xray14 data set with a label of "pulmonary emphysema" actually had subcutaneous emphysema, a failure in the original labeling because the key words were not disambiguated [6]. Tension remains in the quest for representative data sets that are generalizable.

INCOMPLETE EXAMINATION OF FALSE-POSITIVES AND FALSE-NEGATIVES

Another lack of reported failure of AI models until now has been the missing algorithmic audits. The algorithmic audit should be reported as an in-depth examination of false-positives and false-negatives. This is difficult and expensive. A radiologist and a machine learning engineer must be brought together to read cases where the AI failed and the language of communication between the two disciplines requires translation. An audit of an AI algorithm to detect pneumothorax on chest x-rays showed high algorithm performance when Mach bands and chest tubes were present and lower performance when no chest tubes were present [7]. In other words, the algorithm learned that chest tubes predict pneumothoraxes. This is clinically significant because missing an untreated pneumothorax (ie, without a chest tube) may result in adverse patient outcomes. When the radiologist and the machine learning engineer communicate to identify such patterns, the algorithm can be retrained to ignore these findings, and the algorithm performance can be improved. Such insight can only be obtained by diving deep into failures and attempting to decipher why an algorithm fails.

SUCCESSFUL REPORTING OF FAILURES

Not all failures of AI models are obscured by researchers in their articles. Several recent articles have shown that much work is required in the generalizability of AI models. Yi et al showed that two first-year radiology residents were able to outperform a deep learning model to detect pneumothorax from chest x-rays [8]. The AI model used in this study only had an AUC of 0.841. The 2019 SIIM Pneumothorax segmentation competition showcased better models that were more accurate on the same data set but still not generalizable to other data sets. Notably, in competitions such as these, overfitting to the train and test data sets are deliberate practices by participants to win but result in models that are not generalizable and thus cannot be used in the real world.

We hope that embedded failures in AI models will be uncovered through the growth of multi-institutional, multimodality, and accessible data sets. Increased work in human-centered explainable AI, that is, AI research that acknowledges the interplay of human values, clinical practice standards, and AI, is encouraging. This interplay is necessary to amplify trust in AI models and to unpack and mitigate risks in the implementation of AI in clinical radiology practice.

ACKNOWLEDGMENTS

Funding support for Dr Gichoya and Dr Saptarshi was received from the US National Science Foundation #1928481 from the Division of Electrical, Communication & Cyber Systems.

REFERENCES

1. Rubin GD, Lyo JK, Paik DS, et al. Pulmonary nodules on multi–detector row CT scans: performance comparison of radiologists and computer-aided detection. Radiology 2005;234:274-83.
2. Rafferty EA, Park JM, Philpotts LE, et al. Assessing radiologist performance using combined digital mammography and breast tomosynthesis compared with digital mammography alone: results of a multicenter, multireader trial. Radiology 2013;266:104-13.
3. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. J Natl Cancer Inst 2019;111:916-22.
4. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for

breast cancer screening. Nature 2020;577:89-94.
5. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging, In Proceedings of the ACM Conference on Health, Inference, and Learning (CHIL '20). Association for Computing Machinery, New York, NY, USA. 2020:151-9. https://doi.org/10.1145/3368555.3384468.
6. Oakden-Rayner L. Exploring large-scale public medical image datasets. Acad Radiol 2020;27:106-12.
7. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med 2018;15:e1002683.
8. Yi PH, Kim TK, Yu AC, Bennett B, Eng J, Lin CT. Can AI outperform a junior resident? Comparison of deep neural network to first-year radiology residents for identification of pneumothorax. Emerg Radiol 2020;27:367-75.

Saptarshi Purkayastha, PhD, Department of BioHealth Informatics, Indiana University-Purdue University Indianapolis, Indianapolis, Indiana. Hari Trivedi, MD, and Judy Wawira Gichoya, MD, MS, are from the Department of Radiology & Imaging Sciences, Emory University, Atlanta, Georgia.

All authors are employees.

Dr Purkayastha and Dr Gichoya report grants from National Science Foundation, during the conduct of the study. The authors have no conflict of interest related to the material discussed in this article.

Saptarshi Purkayastha, PhD: Department of BioHealth Informatics, Indiana University-Purdue University Indianapolis, 719 Indiana Ave, WK 119, Indianapolis, IN 46202; e-mail: saptpurk@iupui.edu