Radiology: Artificial Intelligence

ORIGINAL RESEARCH

External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review

Alice C. Yu, MD • Bahram Mohajer, MD, MPH • John Eng, MD

From the Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University School of Medicine, 1800 Orleans St, Baltimore, MD 21287.

Received February 25, 2021; revision requested April 5; revision received March 9, 2022; accepted April 12. **Address correspondence to** J.E. (email: *jeng@jhmi.edu*).

Authors declared no funding for this work.

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2022; 4(3):e210064 • https://doi.org/10.1148/ryai.210064 • Content code: AI

Purpose: To assess generalizability of published deep learning (DL) algorithms for radiologic diagnosis.

Materials and Methods: In this systematic review, the PubMed database was searched for peer-reviewed studies of DL algorithms for image-based radiologic diagnosis that included external validation, published from January 1, 2015, through April 1, 2021. Studies using nonimaging features or incorporating non-DL methods for feature extraction or classification were excluded. Two reviewers independently evaluated studies for inclusion, and any discrepancies were resolved by consensus. Internal and external performance measures and pertinent study characteristics were extracted, and relationships among these data were examined using nonparametric statistics.

Results: Eighty-three studies reporting 86 algorithms were included. The vast majority (70 of 86, 81%) reported at least some decrease in external performance compared with internal performance, with nearly half (42 of 86, 49%) reporting at least a modest decrease (≥0.05 on the unit scale) and nearly a quarter (21 of 86, 24%) reporting a substantial decrease (≥0.10 on the unit scale). No study characteristics were found to be associated with the difference between internal and external performance.

Conclusion: Among published external validation studies of DL algorithms for image-based radiologic diagnosis, the vast majority demonstrated diminished algorithm performance on the external dataset, with some reporting a substantial performance decrease.

Supplemental material is available for this article.

©RSNA, 2022

Deep learning (DL) algorithms, predominantly employing convolutional neural networks, have been associated with high diagnostic accuracy in a growing number of classification tasks in medical imaging (1–3). Compared with other machine learning methods, DL algorithms have several advantages. DL algorithms have performed with similar, if not higher, accuracy for classifying large imaging datasets (4–6) compared with conventional machine learning methods, such as support vector machines. Furthermore, DL algorithms do not require labor-intensive feature identification and extraction for data reduction. Reported accuracies of DL algorithms are beginning to match or even exceed those of radiologists (1,7,8).

As we consider the potential applications of DL algorithms to radiology practice, we must consider whether these research results are applicable to the general population. Clinical imaging research is particularly challenging to interpret because of selection bias and the reliance on retrospective data sources (9,10). In both clinical and data science research, data represent characteristics that have a certain distribution in the research population. Selection bias occurs when the distribution in the research population is unknowingly different from that of the general population.

In machine learning research, the issue of selection bias has been recognized under alternate names, such as “dataset shift” (11). Diagnostic imaging applications of machine learning algorithms are even more susceptible to this problem because their performance is entirely dependent on the original development data. If an algorithm’s high diagnostic accuracy depends on hidden peculiarities in the development data with respect to patient population, clinical setting, imaging equipment, and distribution of imaging findings, then the algorithm may not perform well in a general, more diverse population (12,13).

Therefore, to assess real-world clinical efficacy, it is essential to know an algorithm’s performance on an external dataset, one derived from a source that is different than the development data and not used in the algorithm’s training. While the importance of considering external validation in artificial intelligence research is increasingly recognized (14,15), it has been performed in relatively few published studies (16).

To gain a better estimation of the generalizability of DL algorithms for image-based radiologic diagnosis, we conducted a systematic review of studies of DL algorithms that employed an external dataset to perform external validation. We sought to obtain an estimate of the magnitude of performance differences on external datasets and to investigate whether basic study characteristics affect external validation results.

Materials and Methods

Literature Search

This study was a systematic review and was therefore exempt from review by our institutional review board. On May 1, 2021, we searched PubMed for studies published

This copy is for personal use only. To order printed copies, contact reprints@rsna.org

Abbreviations

AUC = area under the receiver operating characteristic curve, DL = deep learning

Summary

Published external validation studies of deep learning for radiologic diagnosis are infrequent, with the vast majority reporting diminished performance in the external dataset compared with the dataset used for algorithm development.

Key Points

- Studies of deep learning algorithms for radiologic diagnosis infrequently include an external dataset, with our systematic review identifying 83 published studies that performed external validation over a 6-year period.
- Nearly half of studies that performed external validation reported at least a modest decrease in external performance, with nearly a quarter reporting a substantial decrease.

Keywords

Meta-Analysis, Computer Applications–Detection/Diagnosis, Neural Networks, Computer Applications–General (Informatics), Epidemiology, Technology Assessment, Diagnosis, Informatics

in the English language from January 1, 2015, through April 1, 2021, on DL algorithms for radiologic diagnosis from medical images, using the search phrase shown in Figure 1. We also reviewed the reference lists of relevant articles for eligible studies. We chose a starting date that was 2 years prior to the release of the National Institutes of Health ChestX-ray14 dataset (17) and the conclusion of the Radiological Society of North America Pneumonia Challenge (18). We assumed that studies published prior to these major events were highly unlikely to meet inclusion criteria.

Study Selection

We considered all studies that evaluated DL algorithms for performing diagnostic classification using radiologic images as direct input. We selected only studies that included external validation of the final algorithm using an external data source from a facility or institution different from that used to develop the algorithm.

Our review focused on the task of diagnostic classification to limit heterogeneity of the included studies. Therefore, we excluded studies that involved tasks other than patient-level diagnostic classification (for example, image segmentation, worklist triage). For a similar reason, we also excluded studies that involved nonimaging clinical features (for example, age, biomarkers, genomic data), methods other than DL for either feature extraction or classification (for example, support vector machines), and feature extraction requiring an expert reader (for example, radiomic data). We excluded animal or phantom studies, review articles, and clinical applications outside of radiology.

Three physicians with 19 (J.E.), 4 (B.M.), and 1 (A.C.Y.) years of experience in conducting systematic reviews in radiology independently assessed titles and abstracts to identify potentially relevant articles for inclusion. The full text of potentially relevant articles was reviewed to identify those meeting inclusion criteria, if necessary. Discrepancies between the reviewers were resolved by consensus.

Data Extraction

For each eligible study, one investigator extracted pertinent information from the full text, including classification task characteristics, labeling method, DL architecture, use of validation, dataset characteristics, performance results, and publication characteristics (Table 1). A second investigator reviewed the extracted data for accuracy, and discrepancies were resolved by consensus. A primary performance measure was identified for each study; in order of preference, we looked for area under the receiver operating characteristic curve (AUC), sensitivity and specificity together, or overall accuracy (proportion of cases correctly classified).

For each study, a representative performance difference was defined as the difference between the primary performance measures of the development and external data sources. Clinically conservative choices were made for studies that reported multiple performance measures, such as those involving multiple institutions. For these studies, the greatest absolute difference between development and external sources was chosen as a representative difference for purposes of categorization and statistical analysis. For studies that reported both sensitivity and specificity differences, the more negative difference was chosen as the representative difference for purposes of analysis. For studies involving multiple institutions for either the development or external datasets, size and disease prevalence were averaged for the purposes of analysis.

On the basis of our experience with receiver operating characteristic analysis, the performance differences between the development and external data sources were grouped for convenience. Performance differences were considered “substantial” if the difference was 0.10 or greater on a positive or negative unit scale, “modest” if less than 0.10 but greater than or equal to 0.05, or “little change” if less than 0.05.

Classification task difficulty was captured in two variables: conspicuity of image findings and composition of nondiseased cases. Conspicuity was classified as “major” (can be confidently diagnosed by imaging alone), “subtle” (diagnosis associated with uncertainty, usually requiring tissue sampling), or “imperceptible” (imaging not usually involved in diagnosis). The second variable, composition of nondiseased cases, indicated whether the “negative” cases were all normal or contained diagnoses other than the index diagnosis. As an indicator of reporting quality, we recorded whether each eligible study stated compliance with any published guideline, such as the Checklist for Artificial Intelligence in Medical Imaging (ie, CLAIM) (15).

Statistical Analysis

The main dependent variable was the representative performance difference between the development and external data sources, computed as external performance minus development performance. Relationships between the dependent variable and pertinent study characteristics were evaluated using various statistical tests, depending on variable type. Relationships between the dependent variable and binary categorical covariates, such as CT versus radiography, were explored with the Wilcoxon rank sum (Mann-Whitney *U*) test. Relationships with dataset size or

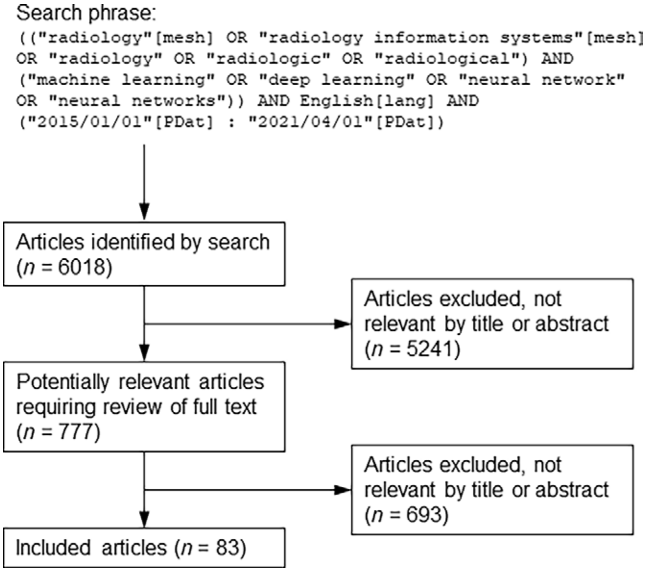
Search phrase:
(("radiology"[mesh] OR "radiology information systems"[mesh] OR "radiology" OR "radiologic" OR "radiological") AND ("machine learning" OR "deep learning" OR "neural network" OR "neural networks")) AND English[lang] AND ("2015/01/01"[PDat] : "2021/04/01"[PDat])
Articles identified by search (n = 6018)
Articles excluded, not relevant by title or abstract (n = 5241)
Potentially relevant articles requiring review of full text (n = 777)
Articles excluded, not relevant by title or abstract (n = 693)
Included articles (n = 83)

Figure 1: Diagram summarizing literature search and article selection.

disease prevalence were explored with the Spearman rank correlation coefficient. The Wilcoxon signed rank test was used to compare paired covariates, such as development versus external dataset sizes. Statistical analysis was performed with the Stata package (version 17; StataCorp). A two-sided P value less than .05 was considered statistically significant.

Results

Search Results

A total of 6018 articles were screened, yielding 83 published articles that met inclusion criteria (2,19–100) (Fig 1). Three of the articles (2,31,68) each reported two major classification tasks being performed by separate algorithms. These three additional tasks were treated as separate studies in the subsequent analysis, resulting in a total of 86 studies.

Study Characteristics

Characteristics of the included studies are shown in Tables 2 and 3. The full table of extracted data is included in Table E1 (supplement). The chest was by far the most common body part imaged for algorithm categorization (41 of 86 studies, 48%), followed by the brain (14 of 86, 16%), bone (10 of 86, 12%), abdomen (seven of 86, 8%), breast (five of 86, 6%), and others (Table E2 [supplement]). Almost three-quarters of studies involved either radiography or CT as the imaging modality. Only three studies implemented prospective data collection for either the development or external dataset, with two of these studies involving diagnosis of COVID-19 (56,94) and one involving thyroid cancer diagnosis (19). The dataset size and disease prevalence varied widely (Table 3). The sizes of the external datasets were statistically significantly smaller than those of the development datasets ($P < .001$, signed rank test). Multiple convolutional neural network architecture types were represented in the included studies, with ResNet being the most common.

Table 1: Main Extracted Data for Each Eligible Study

Item	Value
Task characteristic	
Body part	Chest, brain, bone
Modality	Radiography, CT, MRI, US
Conspicuity of findings	Major, subtle, imperceptible
All normal "negative" cases	Yes, no
Labeling method	NLP, expert reader
Deep learning architecture	ResNet, Inception, VGGNet
Development included validation step	Yes, no
Dataset characteristic (index and external populations)	
Prospective data collection	Yes, no
Population size	Numerical
Proportion of "positive" cases	Numerical
No. of institutions	Numerical
Performance measure (index and external populations)	AUC, sensitivity, specificity
Publication characteristic	
Bibliographic citation	Text
Adherence to quality guideline	STARD, TRIPOD

Note.—Values for body part, modality, labeling method, deep learning architecture, performance measure, and publication characteristics are major examples. AUC = area under the receiver operating characteristic curve, NLP = natural language processing, STARD = Standards for Reporting of Diagnostic Accuracy Studies (102), TRIPOD = Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (103), VGG = Visual Geometry Group.

DL Algorithm External Validation

The median representative performance difference between development and external data sources was –0.046, with a range of –0.60 to 0.13, and 81% (70 of 86) of studies reporting a negative difference (Fig 2). Forty-nine percent of studies (42 of 86) demonstrated at least modestly lower external performance, and 24% of studies (21 of 86) demonstrated substantially lower external performance (Table 4). A few studies reported higher performance with the external dataset than the one for development, including one study showing the AUC increased from 0.84 with the development test set to 0.97 with an external dataset (42).

We found no evidence of relationships between the results of external validation and the study characteristics we examined, using the representative performance difference between the development and external data sources as the measure of external validation. The study characteristics included body part, modality, conspicuity of imaging findings (major vs subtle), composition of negative cases (normal vs other diagnoses), labeling method (direct vs natural language processing), institutional diversity (single vs multiple institutions), population size, disease prevalence, and presence of a validation step during algorithm development.

Study Quality

Only a small number of studies (11 of 86, 13%) stated adherence to a reporting quality guideline. Six used the Nature Research Reporting Summary, a nonspecific guideline for research (101); four used the Standards for Reporting of Diagnostic Accuracy Studies (ie, STARD) (102); and one used the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (ie, TRIPOD) (103).

Discussion

In this systematic review of external validation of DL for radiologic diagnosis, we found that 81% of studies demonstrated at least some diminished performance in external datasets, with nearly half (49%) of studies reporting at least a modest diminution and nearly a quarter (24%) showing a substantial diminution. Balancing accuracy in a study population with that in the general population is a challenge not unique to machine learning research. This issue, also known as generalizability, has long been recognized and studied in clinical trials (104). In clinical trials, assessing generalizability can be done through a number of methods, such as comparing study and target population characteristics and statistical modeling (105–108). However, applying such methods to DL studies is problematic for two main reasons. First, DL studies seldom provide enough demographic or clinical information about the development dataset to allow assessment for potential selection or other bias. Second, the "black box" nature of DL algorithms means that the most important diagnostic features are usually unknown, making it difficult to assess whether these features could be subject to selection or other bias (12).

Among many hundreds of published DL algorithms for radiologic diagnosis, our systematic review identified 83 published articles that reported algorithm performance on an external dataset. This finding corroborates a systematic review performed by Kim et al (16) that found that only 6% of artificial intelligence publications in medical imaging included external validation. Similarly, Yao et al (109) found that only 16 of 155 studies (10%) in their systematic review of DL applications in neuroradiology included external validation, and Nguyen et al (110) found that one in eight studies (13%) in their systematic review of machine learning algorithms distinguishing glioblastoma multiforme from primary central nervous system lymphoma were tested in an external dataset. Potential reasons for the limited number of external validation studies include the difficulty in obtaining an appropriate external dataset

Table 2: Characteristics of Included Studies

Study Characteristic	No. of Studies (*n* = 86)
Body part	
Chest	41 (48)
Not chest	45 (52)
Modality	
Radiography	27 (31)
CT	37 (43)
Other	22 (26)
Conspicuity	
Major	30 (35)
Subtle	45 (52)
Imperceptible	11 (13)
"Negative" cases all normal	
Yes	24 (28)
No	62 (72)
Labeling generated by NLP	
Yes	9 (10)
No	77 (90)
Development included validation step	
Yes	69 (80)
No	17 (20)
Primary performance measure	
AUC	69 (80.2)
Sensitivity and/or specificity	9 (10.5)
Accuracy	5 (5.8)
Free-response AUC	1 (1.1)
F measure	2 (2.3)

Note.—Data in parentheses are percentages. AUC = area under the receiver operating characteristic curve, NLP = natural language processing.

Table 3: Comparison of Development and External Data Sources in Included Studies

Characteristic	Development Data Sources (*n* = 86)	External Data Sources (*n* = 86)
No. of cases		
Median	1167	240
Interquartile range	603–11 455	104–724
Range	25–1 200 000	18–166 578
Prevalence of "positive" diagnosis (%)		
Median	37	47
Interquartile range	23–54	26–53
Range	1–96	1–100
Multi-institutional (%)	44 (38/86)	43 (37/86)

Note.—Data in parentheses are numerator/denominator.

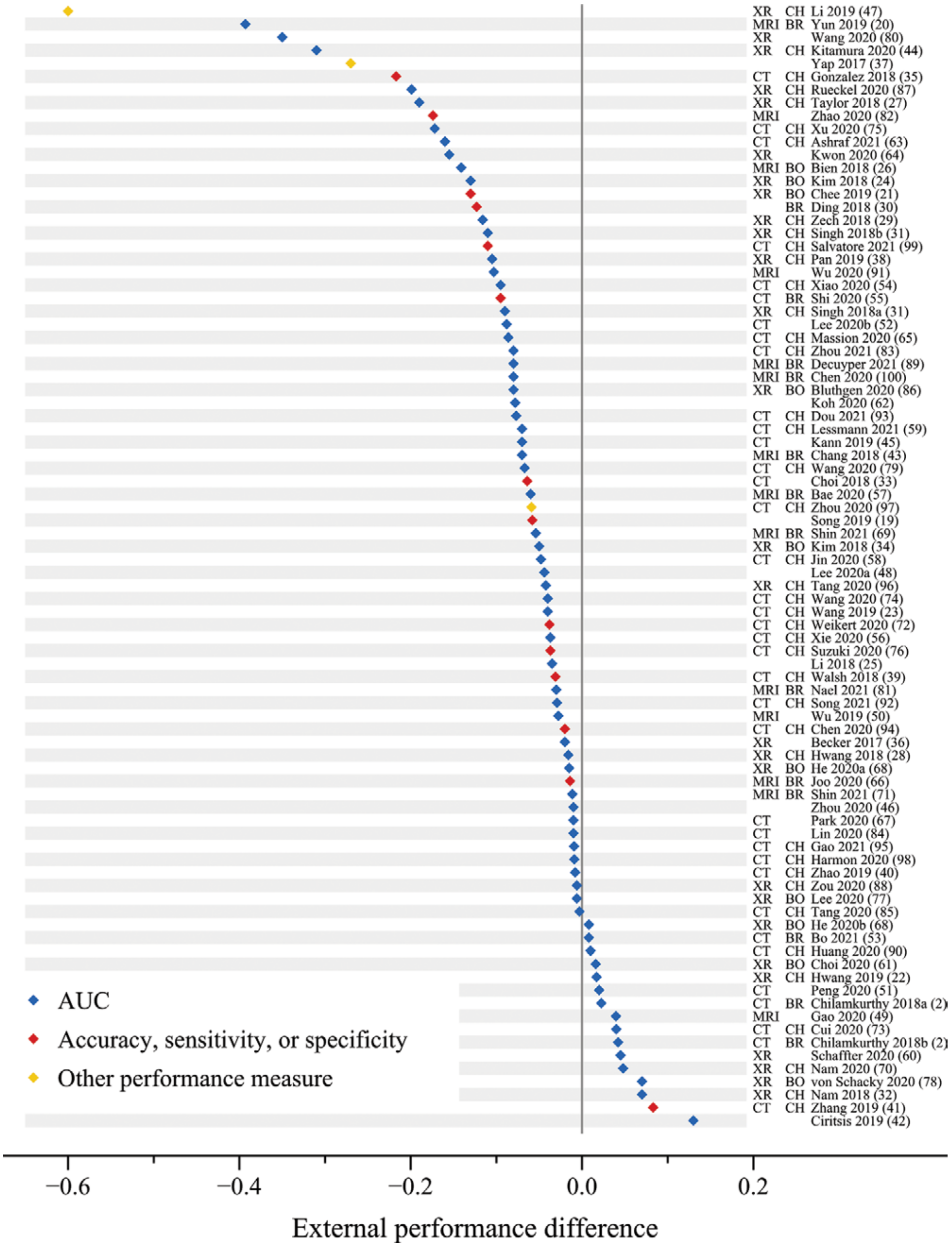

Figure 2: Plot of representative diagnostic performance difference between external and development datasets. The three most common imaging modalities and body parts are indicated. AUC = area under the receiver operating characteristic curve, BO = bone, BR = brain, CH = chest, XR = radiography.

of medical images and lack of awareness of external validation's importance in establishing clinical value. These challenges may diminish as large public datasets become increasingly available and major journals begin supporting guidelines that highlight the importance of performing external validation (15,111,112).

DL algorithms derived from large datasets are expected to have greater generalizability, as larger datasets are more likely to include a broader feature distribution than smaller datasets. Indeed, prior studies of DL algorithms for nonclassification tasks in medical imaging found that larger, multi-institutional development datasets led to improved generalizability (113,114). In contrast, we did not find the size or number of institutions in the development dataset to have a statistically significant impact on external performance, suggesting that other factors may be involved.

An unexpected finding was that a few studies reported higher performance with the external dataset than the one for development. Such a result might be naively interpreted as evidence that some algorithms are highly generalizable, but such a conclusion should be questioned. Because a machine learning algorithm's "knowledge" is exclusively drawn from the development dataset, a generalizable algorithm is expected to have similar, if not slightly lower, external performance compared with internal development performance. Two potential causes of misleadingly high external performance should be considered. First, the external dataset might contain only images with heavily weighted

Table 4: Algorithm Performance in External Dataset Relative to Development Dataset

External Performance vs Internal	No. of Studies (*n* = 86)
Substantial decrease	21 (24.4)
Modest decrease	21 (24.4)
Little change	40 (46.5)
Modest increase	3 (3.5)
Substantial increase	1 (1.1)
Total	86 (100)

Note.—Data in parentheses are percentages.

features responsible for correct classification and not be representative of a realistic target population. Second, the image data might contain information about the diagnosis that is unrelated to the disease process, such as a radiography marker or "burned-in" text in the images. In machine learning, this unintentional information is known as data leakage (115) and is analogous to the epidemiologic concept of a confounding variable. Interpretability techniques such as image embedding and activation maps can help identify data leakage. In the study with the most dramatic external performance increase (42) in our review, the authors found that the external dataset, which was a publicly available breast US dataset, contained very straightforward examples and possibly only contained heavily weighted features.

Limitations

Our systematic review had several limitations. First and most evident was the heterogeneity of the reviewed studies, especially with respect to body part, imaging modality, disease of interest, diagnostic complexity, and performance measures. Heterogeneity in performance measures includes their inherent sources of variation, such as the dependence of sensitivity and specificity on the reader's interpretation threshold. It is reasonable to suspect additional, potentially substantial heterogeneity with respect to imaging equipment, technique, and protocols, as these details were almost always missing from the reviewed studies. Consequently, the overall heterogeneity of included studies precluded quantitative pooling of study results and limited the statistical power of any subgroup comparisons. It is also possible, however, that population and task heterogeneity among medical imaging applications of DL may not be as important as we envision, as many commonly used DL algorithms already originated from tasks outside of medical imaging.

Second, to limit heterogeneity, we focused on a specific type of machine learning and classification task, excluding major areas such as support vector machines, random forests, image segmentation, feature analysis, and image reconstruction. Therefore, our results do not necessarily apply to these other important areas of machine learning. Future systematic reviews should be dedicated to external validation of these algorithm types and radiologic applications.

Third, most of the reviewed studies were focused on technical development and provided little methodological information or clinical description about the datasets and participant populations that were involved, as evidenced by the infrequent use of reporting quality guidelines. Because of this serious limitation in the literature, we were unable to perform a systematic, meaningful assessment of the quality of the reviewed studies and their risk of bias using standardized reporting guidelines (15). The limited methodological and clinical information also reduces the chance of detecting confounding variables associated with dataset and population characteristics. Quality assessment tools like the widely used Quality Assessment of Diagnostic Accuracy Studies 2 (ie, QUADAS-2) (116) may be limited because they are validated for study results derived from a single population, unlike the population comparisons sought in our review. Last, we recognize that our systematic review was subject to potentially large publication bias, likely leading us to overestimate the summary performance of algorithms in external validation studies meeting our selection criteria.

Future Directions

The specific causes of diminished DL algorithm performance on external datasets are largely unknown. Questions remain about what features are actually important for correct diagnosis by machine learning algorithms (117–119), how these features may be biased in datasets, and how external validation is affected. A better understanding of these questions will be necessary before diagnostic machine learning systems achieve routine clinical radiology practice.

We found that substantial improvement is needed in published descriptions of populations from which DL datasets are derived. These improvements are necessary to allow meaningful assessment of study quality and generalizability.

Conclusion

In conclusion, our systematic review found that the vast majority of external validation studies demonstrated diminished algorithm performance on an external dataset, some reporting a substantial performance decrease. Our findings stress the importance of including an external dataset to evaluate the generalizability of DL algorithms, which would improve the quality of future DL studies.

Author contributions: Guarantors of integrity of entire study, **A.C.Y., J.E.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, all authors; clinical studies, **A.C.Y.**; statistical analysis, **B.M., J.E.**; and manuscript editing, all authors

Disclosures of conflicts of interest: **A.C.Y.** No relevant relationships. **B.M.** No relevant relationships. **J.E.** No relevant relationships.

References

1. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 2018;15(11):e1002686.
2. Chilamkurthy S, Ghosh R, Tanamala S, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet 2018;392(10162):2388–2396.

3. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316(22):2402–2410.
4. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017;60(6):84–90.
5. Maruyama T, Hayashi N, Sato Y, et al. Comparison of medical image classification accuracy among three machine learning methods. J XRay Sci Technol 2018;26(6):885–893.
6. Wang H, Zhou Z, Li Y, et al. Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18F-FDG PET/CT images. EJNMMI Res 2017;7(1):11.
7. Wu N, Phang J, Park J, et al. Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans Med Imaging 2020;39(4):1184–1194.
8. Kuo W, Häne C, Mukherjee P, Malik J, Yuh EL. Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning. Proc Natl Acad Sci U S A 2019;116(45):22737–22745.
9. Eng J. Getting started in radiology research: asking the right question and identifying an appropriate study population. Acad Radiol 2004;11(2):149–154.
10. Blackmore CC. The challenge of clinical radiology research. AJR Am J Roentgenol 2001;176(2):327–331.
11. Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. Pattern Recognit 2012;45(1):521–530.
12. Yu AC, Eng J. One algorithm may not fit all: how selection bias affects machine learning performance. RadioGraphics 2020;40(7):1932–1937.
13. Park SH. Diagnostic case-control versus diagnostic cohort studies for clinical validation of artificial intelligence algorithm performance. Radiology 2019;290(1):272–273.
14. Bluemke DA, Moy L, Bredella MA, et al. Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers-from the Radiology editorial board. Radiology 2020;294(3):487–489.
15. Mongan J, Moy L, Kahn CE Jr. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2020;2(2):e200029.
16. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean J Radiol 2019;20(3):405–410.
17. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-Ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, July 21–26, 2017. Piscataway, NJ: IEEE, 2017; 3462–3471.
18. Pan I, Cadrin-Chênevert A, Cheng PM. Tackling the Radiological Society of North America pneumonia detection challenge. AJR Am J Roentgenol 2019;213(3):568–574.
19. Song J, Chai YJ, Masuoka H, et al. Ultrasound image analysis using deep learning algorithm for the diagnosis of thyroid nodules. Medicine (Baltimore) 2019;98(15):e15133.
20. Yun J, Park JE, Lee H, Ham S, Kim N, Kim HS. Radiomic features and multilayer perceptron network classifier: a robust MRI classification strategy for distinguishing glioblastoma from primary central nervous system lymphoma. Sci Rep 2019;9(1):5746.
21. Chee CG, Kim Y, Kang Y, et al. Performance of a deep learning algorithm in detecting osteonecrosis of the femoral head on digital radiography: a comparison with assessments by radiologists. AJR Am J Roentgenol 2019;213(1):155–162.
22. Hwang EJ, Park S, Jin KN, et al. Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on chest radiographs. JAMA Netw Open 2019;2(3):e191095 [Published correction appears in JAMA Netw Open 2019;2(4):e193260.].
23. Wang S, Shi J, Ye Z, et al. Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning. Eur Respir J 2019;53(3):1800986.
24. Kim T, Heo J, Jang DK, et al. Machine learning for detecting moyamoya disease in plain skull radiography using a convolutional neural network. EBioMedicine 2019;40:636–642.
25. Li X, Zhang S, Zhang Q, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. Lancet Oncol 2019;20(2):193–201.
26. Bien N, Rajpurkar P, Ball RL, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. PLoS Med 2018;15(11):e1002699.
27. Taylor AG, Mielke C, Mongan J. Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks: A retrospective study. PLoS Med 2018;15(11):e1002697.
28. Hwang EJ, Park S, Jin KN, et al. Development and validation of a deep learning–based automatic detection algorithm for active pulmonary tuberculosis on chest radiographs. Clin Infect Dis 2019;69(5):739–747.
29. 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(11):e1002683.
30. Ding Y, Sohn JH, Kawczynski MG, et al. A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain. Radiology 2019;290(2):456–464.
31. Singh R, Kalra MK, Nitiwarangkul C, et al. Deep learning in chest radiography: Detection of findings and presence of change. PLoS One 2018;13(10):e0204155.
32. Nam JG, Park S, Hwang EJ, et al. Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. Radiology 2019;290(1):218–228.
33. Choi KJ, Jang JK, Lee SS, et al. Development and validation of a deep learning system for staging liver fibrosis by using contrast agent-enhanced CT images in the liver. Radiology 2018;289(3):688–697.
34. Kim Y, Lee KJ, Sunwoo L, et al. Deep learning in diagnosis of maxillary sinusitis using conventional radiography. Invest Radiol 2019;54(1):7–15.
35. González G, Ash SY, Vegas-Sánchez-Ferrero G, et al. Disease staging and prognosis in smokers using deep learning in chest computed tomography. Am J Respir Crit Care Med 2018;197(2):193–203.
36. Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. Invest Radiol 2017;52(7):434–440.
37. Yap MH, Pons G, Martí J, et al. Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J Biomed Health Inform 2018;22(4):1218–1226.
38. Pan I, Agarwal S, Merck D. Generalizable inter-institutional classification of abnormal chest radiographs using efficient convolutional neural networks. J Digit Imaging 2019;32(5):888–896.
39. Walsh SLF, Calandriello L, Silva M, Sverzellati N. Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. Lancet Respir Med 2018;6(11):837–845.
40. Zhao W, Yang J, Ni B, et al. Toward automatic prediction of EGFR mutation status in pulmonary adenocarcinoma with 3D deep learning. Cancer Med 2019;8(7):3532–3543.
41. Zhang C, Sun X, Dang K, et al. Toward an expert level of lung cancer detection and classification using a deep convolutional neural network. Oncologist 2019;24(9):1159–1165.
42. Ciritsis A, Rossi C, Eberhard M, Marcon M, Becker AS, Boss A. Automatic classification of ultrasound breast lesions using a deep convolutional neural network mimicking human decision-making. Eur Radiol 2019;29(10):5458–5468.
43. Chang K, Bai HX, Zhou H, et al. Residual convolutional neural network for the determination of IDH status in low- and high-grade gliomas from MR imaging. Clin Cancer Res 2018;24(5):1073–1081.
44. Kitamura G, Deible C. Retraining an open-source pneumothorax detecting machine learning algorithm for improved performance to medical images. Clin Imaging 2020;61:15–19.
45. Kann BH, Hicks DF, Payabvash S, et al. Multi-institutional validation of deep learning for pretreatment identification of extranodal extension in head and neck squamous cell carcinoma. J Clin Oncol 2020;38(12): 1304–1311.
46. Zhou LQ, Wu XL, Huang SY, et al. Lymph node metastasis prediction from primary breast cancer US images using deep learning. Radiology 2020;294(1):19–28.
47. Li X, Shen L, Xie X, et al. Multi-resolution convolutional networks for chest X-ray radiograph based lung nodule detection. Artif Intell Med 2020;103:101744.
48. Lee JH, Joo I, Kang TW, et al. Deep learning with ultrasonography: automated classification of liver fibrosis using a deep convolutional neural network. Eur Radiol 2020;30(2):1264–1273.
49. Gao X, Wang X. Performance of deep learning for differentiating pancreatic diseases on contrast-enhanced magnetic resonance imaging: A preliminary study. Diagn Interv Imaging 2020;101(2):91–100.
50. Wu J, Xin J, Yang X, et al. Deep morphology aided diagnosis network for segmentation of carotid artery vessel wall and diagnosis of carotid atherosclerosis on black-blood vessel wall MRI. Med Phys 2019;46(12):5544–5561. [Published correction appears in Med Phys 2020;47(6):2575.]

51. Peng J, Kang S, Ning Z, et al. Residual convolutional neural network for predicting response of transarterial chemoembolization in hepatocellular carcinoma from CT imaging. Eur Radiol 2020;30(1):413–424.
52. Lee JH, Ha EJ, Kim D, et al. Application of deep learning to the diagnosis of cervical lymph node metastasis from thyroid cancer with CT: external validation and clinical utility for resident training. Eur Radiol 2020;30(6):3066–3072.
53. Bo ZH, Qiao H, Tian C, et al. Toward human intervention-free clinical diagnosis of intracranial aneurysm via deep neural network. Patterns (N Y) 2021;2(2):100197.
54. Xiao LS, Li P, Sun F, et al. Development and validation of a deep learning-based model using computed tomography imaging for predicting disease severity of coronavirus disease 2019. Front Bioeng Biotechnol 2020;8:898.
55. Shi Z, Miao C, Schoepf UJ, et al. A clinically applicable deep-learning model for detecting intracranial aneurysm in computed tomography angiography images. Nat Commun 2020;11(1):6090.
56. Xie Q, Lu Y, Xie X, et al. The usage of deep neural network improves distinguishing COVID-19 from other suspected viral pneumonia by clinicians on chest CT: a real-world study. Eur Radiol 2021;31(6):3864–3873.
57. Bae JB, Lee S, Jung W, et al. Identification of Alzheimer's disease using a convolutional neural network model based on T1-weighted magnetic resonance imaging. Sci Rep 2020;10(1):22252.
58. Jin C, Chen W, Cao Y, et al. Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. Nat Commun 2020;11(1):5088.
59. Lessmann N, Sánchez CI, Beenen L, et al. Automated assessment of COVID-19 Reporting and Data System and chest CT severity scores in patients suspected of having COVID-19 using artificial intelligence. Radiology 2021;298(1):E18–E28.
60. Schaffter T, Buist DSM, Lee CI, et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. JAMA Netw Open 2020;3(3):e200265.
61. Choi JW, Cho YJ, Lee S, et al. Using a dual-input convolutional neural network for automated detection of pediatric supracondylar fracture on conventional radiography. Invest Radiol 2020;55(2):101–110.
62. Koh J, Lee E, Han K, et al. Diagnosis of thyroid nodules on ultrasonography by a deep convolutional neural network. Sci Rep 2020;10(1):15245.
63. Ashraf SF, Yin K, Meng CX, et al. Predicting benign, preinvasive, and invasive lung nodules on computed tomography scans using machine learning. J Thorac Cardiovasc Surg 2022;163(4):1496–1505.e10.
64. Kwon G, Ryu J, Oh J, et al. Deep learning algorithms for detecting and visualising intussusception on plain abdominal radiography in children: a retrospective multicenter study. Sci Rep 2020;10(1):17582.
65. Massion PP, Antic S, Ather S, et al. Assessing the accuracy of a deep learning method to risk stratify indeterminate pulmonary nodules. Am J Respir Crit Care Med 2020;202(2):241–249.
66. Joo B, Ahn SS, Yoon PH, et al. A deep learning algorithm may automate intracranial aneurysm detection on MR angiography with high diagnostic performance. Eur Radiol 2020;30(11):5785–5793.
67. Park JJ, Kim KA, Nam Y, Choi MH, Choi SY, Rhie J. Convolutional-neural-network-based diagnosis of appendicitis via CT scans in patients with acute abdominal pain presenting in the emergency department. Sci Rep 2020;10(1):9556.
68. He Y, Pan I, Bao B, et al. Deep learning-based classification of primary bone tumors on radiographs: A preliminary study. EBioMedicine 2020;62:103121.
69. Shin I, Kim H, Ahn SS, et al. Development and validation of a deep learning-based model to distinguish glioblastoma from solitary brain metastasis using conventional MR images. AJNR Am J Neuroradiol 2021;42(5):838–844.
70. Nam JG, Kim M, Park J, et al. Development and validation of a deep learning algorithm detecting 10 common abnormalities on chest radiographs. Eur Respir J 2021;57(5):2003061.
71. Shin DH, Heo H, Song S, et al. Automated assessment of the substantia nigra on susceptibility map-weighted imaging using deep convolutional neural networks for diagnosis of Idiopathic Parkinson's disease. Parkinsonism Relat Disord 2021;85:84–90.
72. Weikert T, Noordtzij LA, Bremerich J, et al. Assessment of a deep learning algorithm for the detection of rib fractures on whole-body trauma computed tomography. Korean J Radiol 2020;21(7):891–899.
73. Cui S, Ming S, Lin Y, et al. Development and clinical application of deep learning model for lung nodules screening on CT images. Sci Rep 2020;10(1):13657.
74. Wang S, Zha Y, Li W, et al. A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis. Eur Respir J 2020;56(2):2000775.
75. Xu X, Wang C, Guo J, et al. MSCS-DeepLN: Evaluating lung nodule malignancy using multi-scale cost-sensitive neural networks. Med Image Anal 2020;65:101772.
76. Suzuki K, Otsuka Y, Nomura Y, Kumamaru KK, Kuwatsuru R, Aoki S. Development and validation of a modified three-dimensional U-Net deep-learning model for automated detection of lung nodules on chest ct images from the lung image database consortium and Japanese datasets. Acad Radiol 2022;29(Suppl 2):S11–S17.
77. Lee KJ, Ryoo I, Choi D, Sunwoo L, You SH, Jung HN. Performance of deep learning to detect mastoiditis using multiple conventional radiographs of mastoid. PLoS One 2020;15(11):e0241796.
78. von Schacky CE, Sohn JH, Liu F, et al. Development and validation of a multitask deep learning model for severity grading of hip osteoarthritis features on radiographs. Radiology 2020;295(1):136–145.
79. Wang M, Xia C, Huang L, et al. Deep learning-based triage and analysis of lesion burden for COVID-19: a retrospective study with external validation. Lancet Digit Health 2020;2(10):e506–e515.
80. Wang X, Liang G, Zhang Y, Blanton H, Bessinger Z, Jacobs N. Inconsistent performance of deep learning models on mammogram classification. J Am Coll Radiol 2020;17(6):796–803.
81. Nael K, Gibson E, Yang C, et al. Automated detection of critical findings in multi-parametric brain MRI using a system of 3D neural networks. Sci Rep 2021;11(1):6876.
82. Zhao X, Xie P, Wang M, et al. Deep learning-based fully automated detection and segmentation of lymph nodes on multiparametric-mri for rectal cancer: A multicentre study. EBioMedicine 2020;56:102780.
83. Zhou M, Yang D, Chen Y, et al. Deep learning for differentiating novel coronavirus pneumonia and influenza pneumonia. Ann Transl Med 2021;9(2):111.
84. Lin F, Ma C, Xu J, et al. A CT-based deep learning model for predicting the nuclear grade of clear cell renal cell carcinoma. Eur J Radiol 2020;129:109079.
85. Tang LYW, Coxson HO, Lam S, Leipsic J, Tam RC, Sin DD. Towards large-scale case-finding: training and validation of residual networks for detection of chronic obstructive pulmonary disease using low-dose CT. Lancet Digit Health 2020;2(5):e259–e267.
86. Blüthgen C, Becker AS, Vittoria de Martini I, Meier A, Martini K, Frauenfelder T. Detection and localization of distal radius fractures: deep learning system versus radiologists. Eur J Radiol 2020;126:108925.
87. Rueckel J, Kunz WG, Hoppe BF, et al. Artificial intelligence algorithm detecting lung infection in supine chest radiographs of critically ill patients with a diagnostic accuracy similar to board-certified radiologists. Crit Care Med 2020;48(7):e574–e583.
88. Zou XL, Ren Y, Feng DY, et al. A promising approach for screening pulmonary hypertension based on frontal chest radiographs using deep learning: a retrospective study. PLoS One 2020;15(7):e0236378.
89. Decuyper M, Bonte S, Deblaere K, Van Holen R. Automated MRI based pipeline for segmentation and prediction of grade, IDH mutation and 1p19q co-deletion in glioma. Comput Med Imaging Graph 2021;88:101831.
90. Huang SC, Kothari T, Banerjee I, et al. PENet-a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging. NPJ Digit Med 2020;3(1):61. [Published correction appears in NPJ Digit Med 2020;3:102.]
91. Wu Q, Wang S, Zhang S, et al. Development of a deep learning model to identify lymph node metastasis on magnetic resonance imaging in patients with cervical cancer. JAMA Netw Open 2020;3(7):e2011625.
92. Song Z, Liu T, Shi L, et al. The deep learning model combining CT image and clinicopathological information for predicting ALK fusion status and response to ALK-TKI therapy in non-small cell lung cancer patients. Eur J Nucl Med Mol Imaging 2021;48(2):361–371.
93. Dou Q, So TY, Jiang M, et al. Federated deep learning for detecting COVID-19 lung abnormalities in CT: a privacy-preserving multinational validation study. NPJ Digit Med 2021;4(1):60.
94. Chen J, Wu L, Zhang J, et al. Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography. Sci Rep 2020;10(1):19196.
95. Gao K, Su J, Jiang Z, et al. Dual-branch combination network (DCN): towards accurate diagnosis and lesion segmentation of COVID-19 using CT images. Med Image Anal 2021;67:101836.
96. Tang YX, Tang YB, Peng Y, et al. Automated abnormality classification of chest radiographs using deep convolutional neural networks. NPJ Digit Med 2020;3(1):70.
97. Zhou QQ, Tang W, Wang J, et al. Automatic detection and classification of rib fractures based on patients' CT images and clinical information via convolutional neural network. Eur Radiol 2021;31(6):3815–3825.
98. Harmon SA, Sanford TH, Xu S, et al. Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. Nat Commun 2020;11(1):4080.

99. Salvatore C, Interlenghi M, Monti CB, et al. Artificial intelligence applied to chest X-ray for differential diagnosis of COVID-19 pneumonia. Diagnostics (Basel) 2021;11(3):530.
100. Chen X, Fan Z, Li KKW, et al. Molecular subgrouping of medulloblastoma based on few-shot learning of multitasking using conventional MR images: a retrospective multicenter study. Neurooncol Adv 2020;2(1):vdaa079.
101. Nature Portfolio. Reporting standards and availability of data, materials, code and protocols. https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards. Accessed January 25, 2022.
102. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. Radiology 2015;277(3):826–832.
103. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162(1):55–63.
104. Kramer MS, Shapiro SH. Scientific challenges in the application of randomized trials. JAMA 1984;252(19):2739–2745.
105. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. Am J Epidemiol 2010;172(1):107–115.
106. Anderson TS, Odden M, Penko J, Kazi DS, Bellows BK, Bibbins-Domingo K. Generalizability of clinical trials supporting the 2017 American College of Cardiology/American Heart Association blood pressure guideline. JAMA Intern Med 2020;180(5):795–797.
107. Rothwell PM. Factors that can affect the external validity of randomised controlled trials. PLoS Clin Trials 2006;1(1):e9.
108. Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. Trials 2015;16(1):495.
109. Yao AD, Cheng DL, Pan I, Kitamura F. Deep learning in neuroradiology: a systematic review of current algorithms and approaches for the new wave of imaging technology. Radiol Artif Intell 2020;2(2):e190026.
110. Nguyen AV, Blears EE, Ross E, Lall RR, Ortega-Barnett J. Machine learning applications for the differentiation of primary central nervous system lymphoma from glioblastoma on imaging: a systematic review and meta-analysis. Neurosurg Focus 2018;45(5):E5.
111. England JR, Cheng PM. Artificial intelligence for medical image analysis: a guide for authors and reviewers. AJR Am J Roentgenol 2019;212(3):513–519.
112. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology 2018;286(3):800–809.
113. Remedios SW, Roy S, Bermudez C, et al. Distributed deep learning across multisite datasets for generalized CT hemorrhage segmentation. Med Phys 2020;47(1):89–98.
114. Balki I, Amirabadi A, Levman J, et al. Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. Can Assoc Radiol J 2019;70(4):344–353.
115. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: formulation, detection, and avoidance. ACM Trans Knowl Discov Data 2012;6(4):1–21.
116. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155(8):529–536.
117. AI diagnostics need attention. Nature 2018;555(7696):285.
118. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1(5):206–215.
119. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. Lancet Digit Health 2020;2(9):e489–e492.