International Journal of Computer Assisted Radiology and Surgery
https://doi.org/10.1007/s11548-024-03136-9

ORIGINAL ARTICLE

Performance changes due to differences among annotating radiologists for training data in computerized lesion detection

Yukihiro Nomura[1,2] · Shouhei Hanaoka[3,4] · Naoto Hayashi[2] · Takeharu Yoshikawa[2] · Saori Koshino[3] · Chiaki Sato[5] · Momoko Tatsuta[6] · Yuya Tanaka[4] · Shintaro Kano[3] · Moto Nakaya[4] · Shohei Inui[3] · Masashi Kusakabe[7] · Takahiro Nakao[2] · Soichiro Miki[2] · Takeyuki Watadani[3,4] · Ryusuke Nakaoka[8] · Akinobu Shimizu[9] · Osamu Abe[3,4]

Received: 4 January 2024 / Accepted: 28 March 2024

Abstract

Purpose The quality and bias of annotations by annotators (e.g., radiologists) affect the performance changes in computer-aided detection (CAD) software using machine learning. We hypothesized that the difference in the years of experience in image interpretation among radiologists contributes to annotation variability. In this study, we focused on how the performance of CAD software changes with retraining by incorporating cases annotated by radiologists with varying experience.

Methods We used two types of CAD software for lung nodule detection in chest computed tomography images and cerebral aneurysm detection in magnetic resonance angiography images. Twelve radiologists with different years of experience independently annotated the lesions, and the performance changes were investigated by repeating the retraining of the CAD software twice, with the addition of cases annotated by each radiologist. Additionally, we investigated the effects of retraining using integrated annotations from multiple radiologists.

Results The performance of the CAD software after retraining differed among annotating radiologists. In some cases, the performance was degraded compared to that of the initial software. Retraining using integrated annotations showed different performance trends depending on the target CAD software, notably in cerebral aneurysm detection, where the performance decreased compared to using annotations from a single radiologist.

Conclusions Although the performance of the CAD software after retraining varied among the annotating radiologists, no direct correlation with their experience was found. The performance trends differed according to the type of CAD software used when integrated annotations from multiple radiologists were used.

Keywords Computer-aided detection (CAD) · Machine learning · Retraining · Annotation

Introduction

Computer-aided detection (CAD) software has been developed by numerous research groups, and CAD software using machine learning, particularly deep learning, has increased in recent years [1–5]. The development of CAD software

✉ Yukihiro Nomura
ynomura@chiba-u.jp

[1] Center for Frontier Medical Engineering, Chiba University, 1-33 Yayoi-cho, Inage-ku, Chiba 263-8522, Japan

[2] Department of Computational Diagnostic Radiology and Preventive Medicine, The University of Tokyo Hospital, Tokyo, Japan

[3] Department of Radiology, The University of Tokyo Hospital, Tokyo, Japan

[4] Division of Radiology and Biomedical Engineering, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

[5] Department of Radiology, Tokyo Metropolitan Bokutoh Hospital, Tokyo, Japan

[6] Department of Diagnostic Radiology, Kitasato University Hospital, Sagamihara, Kanagawa, Japan

[7] Department of Radiology, NTT Medical Center Tokyo, Tokyo, Japan

[8] Division of Medical Devices, National Institute of Health Sciences, Kawasaki, Kanagawa, Japan

[9] Institute of Engineering, Tokyo University of Agriculture and Technology, Tokyo, Japan

Published online: 16 April 2024

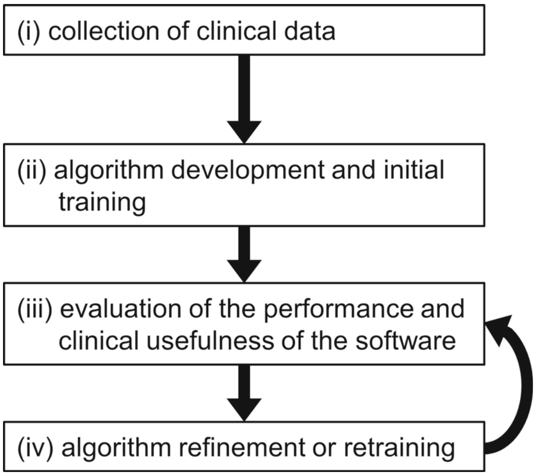

(i) collection of clinical data
(ii) algorithm development and initial training
(iii) evaluation of the performance and clinical usefulness of the software
(iv) algorithm refinement or retraining

Fig. 1 Process for development of CAD software

based on machine learning involves: (i) the collection of clinical data, (ii) algorithm development and initial training, (iii) the evaluation of the performance and clinical usefulness of the software, and (iv) algorithm refinement or retraining (Fig. 1) [6]. This development is not a one-time process, even for commercially available CAD software, and it involves repeated cycles of steps (iii) and (iv). The performance of the CAD software depends on the quality and quantity of the datasets used for machine learning. If the data characteristics differ between development and practical use, the performance of the CAD software degrades. The main factors causing changes in the performance of CAD software are as follows.

(1) Difference in subject populations
(2) Difference in scanners or scan parameters
(3) Quality and bias of annotation by annotators (e.g., radiologists)

Several research groups, including ours, have reported changes in CAD software performance caused by the first two factors [7–14]. These changes can be overcome through continuous data collection and retraining [9, 10, 14]. Regarding Factor (3), for example, in the case of the Lung Image Database Consortium–Image Database Resource Initiative (LIDC-IDRI) database of lung nodules in chest computed tomography (CT) images [15], the diagnoses and defined contours vary among annotating radiologists [16]. Tachibana et al. reported that incorporating information about radiologists who performed annotations in the classification of brain aneurysms in magnetic resonance angiography (MRA) images improved classification performance [17]. Figure 2 shows the examples of annotation variability in spherical regions of interest (ROIs) by two radiologists. These annotations encompassed the entire lesion in three dimensions (3D). We hypothesized that a potential factor contributing to this variability is the difference in years of experience among radiologists in image interpretation. Moreover, integrating annotations from multiple radiologists may reduce variability among annotators. However, to the best of our knowledge, no studies have focused on either the years of experience in image interpretation among annotators or the relationship between methods of annotation integration and the performance of CAD software.

In this study, we investigated the following two items regarding the change in performance in the retraining of CAD software owing to the years of experience in image interpretation among radiologists who perform annotation work.

- Whether a relationship was observed between the years of experience of the annotating radiologists and the performance of the trained CAD software.
- Whether the performance of the CAD software is improved by integrating annotations from multiple radiologists with different years of experience.

We targeted two types of CAD software for lung nodule detection in chest CT images and cerebral aneurysm detection in MRA images, without altering Factors (1) and (2) mentioned above.

Materials and methods

Datasets

This retrospective study was approved by the Ethics Review Board of our institution. We utilized a chest CT dataset for lung nodule detection (Fig. 3a) and a brain MRA dataset for cerebral aneurysm detection (Fig. 3b) collected from our institution. The subjects were adults who underwent annual whole-body general medical examinations, including chest CT or brain MRA. Written informed consent was obtained from all participants to use their clinical images for research. Each dataset consisted of subsets for the initial training, two sets for retraining (Retraining1 and Retraining2), and a test, as shown in Fig. 4. Table 1 shows the number of cases in the chest CT and brain MRA datasets that are common to both datasets. The criteria for selecting the positive and negative cases are described in the following subsections. In the retraining and test sets, 10 ambiguous cases were used as negative cases.

Chest CT dataset

We used a total of 300 cases of chest CT images. The datasets were acquired from a GE LightSpeed CT scanner (GE Healthcare, Waukesha, WI, USA). The original voxel

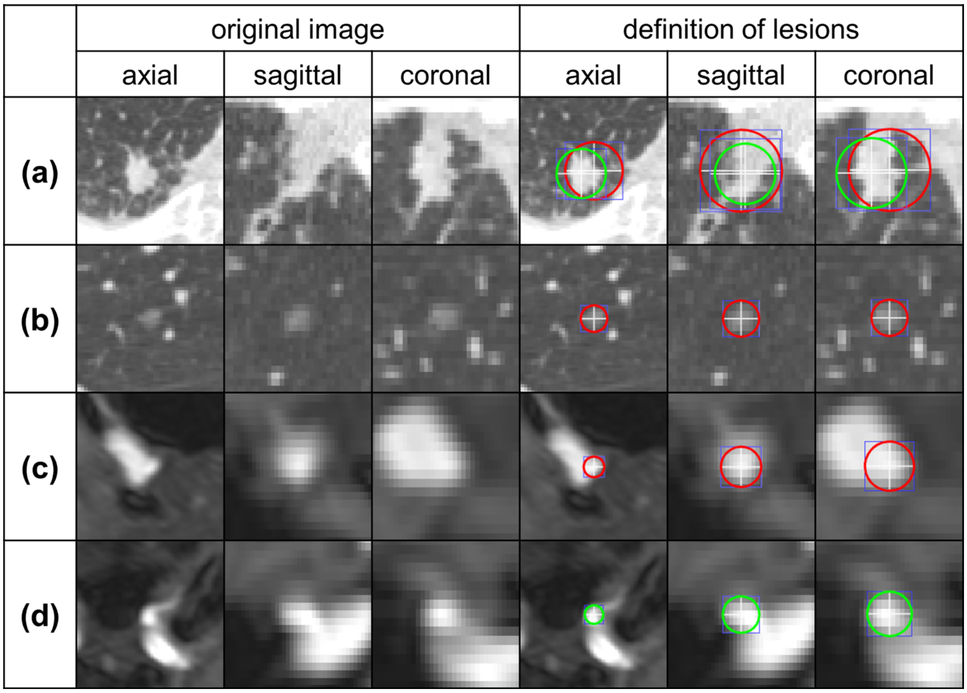


Fig. 2 Examples of mismatch in annotations of spherical ROI by two radiologists. **a** 15.8 mm solid nodule, annotated by both, but with differing ranges. **b** 6.6 mm pure ground-glass nodule, annotated by only one radiologist. **c** 3.2 mm saccular aneurysm, annotated by only one radiologist. **d** Infundibular dilation at the origin of the left ophthalmic artery, incorrectly annotated as a saccular aneurysm by one radiologist

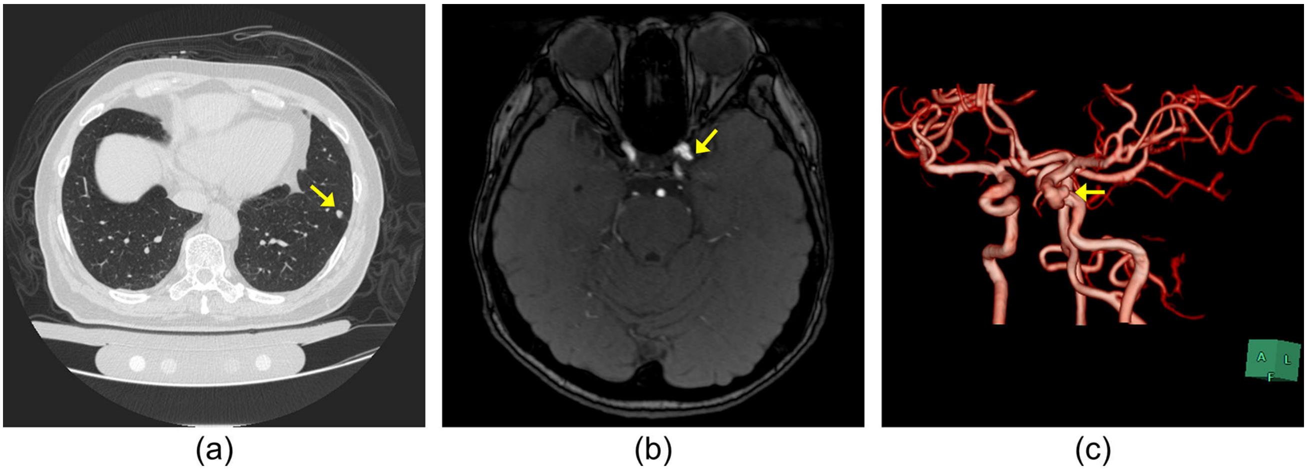
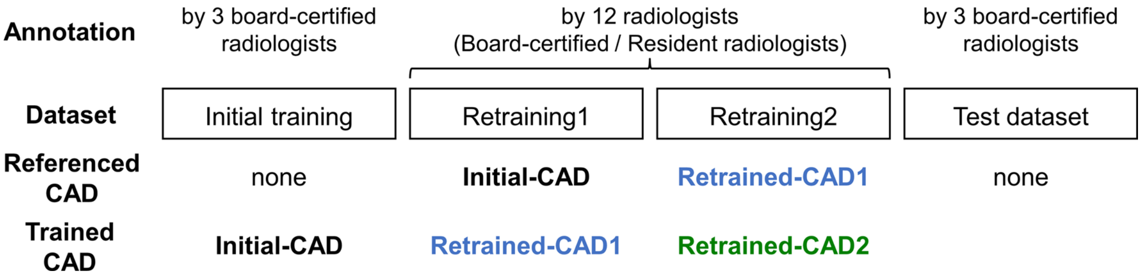


Fig. 3 Examples of target lesions included in the test set. **a** Lung nodule (8.9 mm solid nodule) in chest CT images, **b** cerebral aneurysm (4.3 mm in left internal carotid artery) in brain MRA images, **c** pre-rendered whole-brain volume rendering images of the same case as in (**b**). A yellow arrow indicates the target lesion, but the arrows are not visible during actual reading



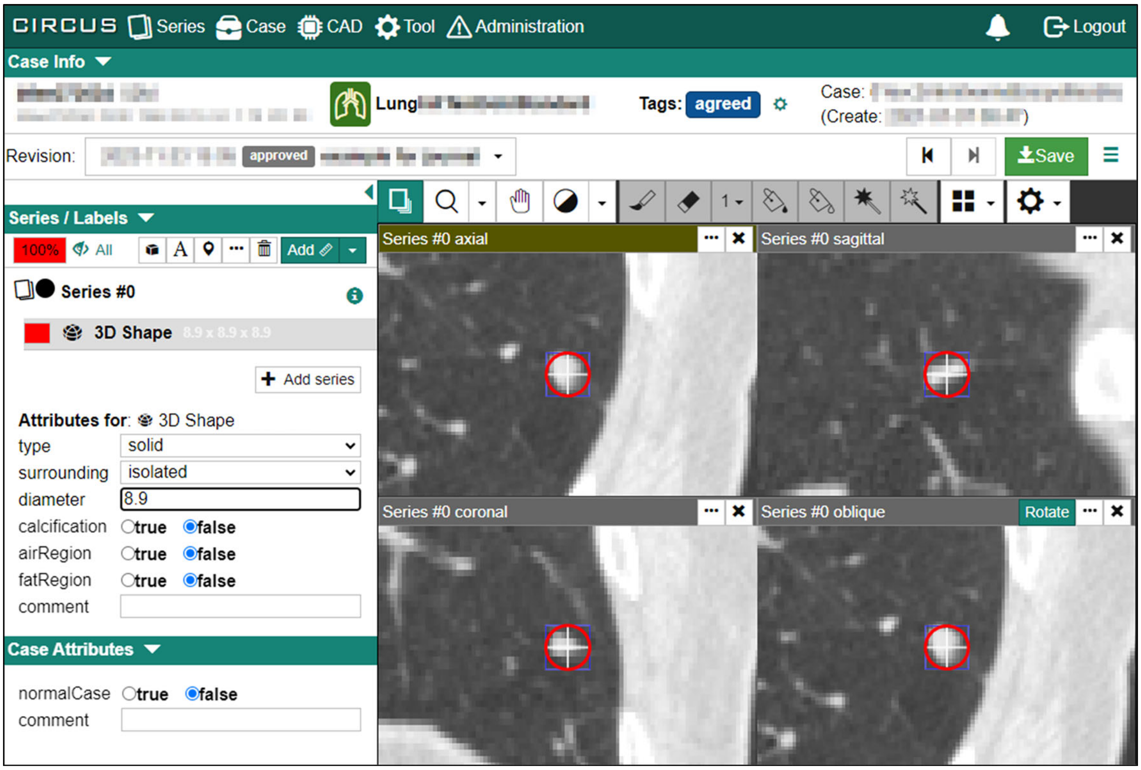
Annotation	by 3 board-certified radiologists	by 12 radiologists (Board-certified / Resident radiologists)		by 3 board-certified radiologists
Dataset	Initial training	Retraining1	Retraining2	Test dataset
Referenced CAD	none	**Initial-CAD**	**Retrained-CAD1**	none
Trained CAD	**Initial-CAD**	**Retrained-CAD1**	**Retrained-CAD2**	

Fig. 4 Procedure of annotation of dataset and training for each CAD software

Table 1 Number of cases for chest CT and brain MRA datasets (common to both datasets)

	Initial training	Retraining1	Retraining2	Test
Positive cases	50	40	40	50
Negative cases	0	35*	35*	50*
Total	50	75	75	100

*Ten cases included lung nodules < 6 mm in diameter in the chest CT dataset and infundibular dilation or basilar artery bifurcation in the brain MRA dataset



Attributes for: 3D Shape

type: solid

surrounding: isolated

diameter: 8.9

calcification ○true ●false

airRegion ○true ●false

fatRegion ○true ●false

comment

normalCase ○true ●false

comment

Fig. 5 Web interface for defining spherical ROI in CIRCUS DB. The left panel has a series selector and an attribute editor. The right panel has a grid of DICOM viewer components, which includes an axial view, a sagittal view, a coronal view, and an oblique view

size was 0.781 × 0.781 × 1.250 mm^3. The acquisition parameters were as follows: number of detector rows, 16; tube voltage, 120 kVp; tube current, 50–290 mA (automatic exposure control); noise index, 20.41; rotation time, 0.5 s; moving table speed, 70 mm/s; body filter, standard; reconstruction slice thickness and interval, 1.25 mm; field of view, 400 mm; matrix size, 512 × 512 pixels; pixel spacing, 0.781 mm. Each positive case included at least one lung nodule with a diameter of 6 mm or more. Two board-certified radiologists (N.H. and S.H., with 32 and 19 years of experience in chest CT interpretation, respectively) annotated the cases in the initial training and test subsets via spherical ROIs to encompass the entire nodule in 3D using the web-based image database system CIRCUS DB [6] (Fig. 5). Discrepancies between the two radiologists were resolved by a third board-certified radiologist (T.W., with 20 years of experience). The cases for retraining subsets were selected based on radiology reports by consensual reading by two experienced radiologists. Ten negative cases in the retraining and test subsets included pulmonary nodules of < 6 mm in diameter.

Brain MRA dataset

We used a total of 300 cases of brain MRA images. These images were acquired using four 3-Tesla MR scanners (two Signa HDxt, GE Healthcare, Waukesha, WI, USA; one Discovery MR750, GE Healthcare; and one Skyra, Siemens Healthcare, Erlangen, Germany). Table 2 presents the details of the examination. Each positive case included at least one saccular aneurysm with a diameter of 2 mm or more. Two board-certified radiologists (N.H. and S.M., with 32 and 14 years of experience in MRA interpretation, respectively) annotated the cases in the initial training and test subsets via spherical ROIs to encompass the entire aneurysm in 3D using

Table 2 Specification of the brain MRA datasets

MR scanners	Scan parameters
Two Signa HDxt and one Discovery MR750 (GE Healthcare, Waukesha, WI, USA)	Field of view (FOV), 240 mm; matrix size, 512 × 512 pixels; pixel spacing, 0.469 mm; slice thickness, 1.2 mm; slice interval, 0.6 mm; repetition time (TR), 22 or 25 ms; echo time (TE), 2.7–3.3 ms; flip angle, 15°
Skyra (Siemens Healthcare, Erlangen, Germany)	FOV, 230 mm; percent phase field of view, 82.3%; matrix size, 768 × 632 pixels; pixel spacing, 0.299 mm; slice thickness, 0.6 mm; slice interval, 0.6 mm; TR, 20 ms; TE, 3.69 ms; flip angle, 13°

the CIRCUS DB. Discrepancies between the two radiologists were resolved by a third board-certified radiologist (T.Y., with 26 years of experience). The cases for the retraining subsets were selected based on radiology reports by consensual reading by two experienced radiologists. Ten of the negative cases in the retraining and test subsets included infundibular dilation (Fig. 1d) or basilar artery bifurcation.

CAD algorithm

Lung nodule detection in CT images

Chest CT images were resampled to a 1.0 mm isotropic voxel size using tricubic interpolation, and the lung volume was extracted using the method described in [18]. Nodule candidates were extracted using volumetric curvature-based thresholding and region growing [19]. Subsequently, for each nodule candidate, a 32 × 32 × 32 cubic volume of interest (VOI) was extracted around the center of gravity of the nodule candidate. The VOIs were fed into a classifier based on 3D ResNet-18 [20] to classify true nodules and false positives (FPs). The hyperparameters of the model were as follows: loss function, cross-entropy loss; optimizer, momentum stochastic gradient descent; learning rate, 1.0×10^{-5}; momentum, 0.99; weight decay, 0.001; minibatch size, 8; number of epochs, 500. The numbers of negative (FP) and positive (true lung nodule) VOIs were equalized using data augmentation and undersampling to address the interclass imbalance in the training data. For each positive VOI, 29 augmented VOIs were generated by random shifts within ± 4 voxels on the x-, y-, and z-axes, random scaling in the range of [0.85, 1.15], and random rotation (0°/90°/180°/270°) in each of the axial, coronal, and sagittal planes. By contrast, negative VOIs were randomly undersampled such that the numbers of negative and positive VOIs were the same. Augmented positive VOIs and sampled negative VOIs were changed for each epoch.

Cerebral aneurysm detection in MRA images

MRA images were resampled to a 0.469 mm isotropic voxel size using tricubic interpolation, and the signal intensity distributions of the images were standardized by global piecewise linear mapping [21]. The arterial region was extracted using the region growing-based method described in [22]. The voxel-based classifier based on the convolutional neural network (CNN) was employed at the voxels of the arterial region. The inputs of the CNN model were two-dimensional images, which were generated from a VOI around the target voxel by applying a maximum intensity projection algorithm. The CNN model consisted of two convolutional layers, two max-pooling layers, and two fully connected layers. The output layer was a single unit, and a logistic function was applied to the output to convert it into a positive probability (ranging from 0 to 1). The hyperparameters of the CNN model were set according to a random search, as previously reported [23]. The number of epochs was set to 10.

Image annotation and retraining CAD

The procedure for annotation and retraining of each CAD software is shown (Fig. 4).

(1) Initial-CAD is the CAD software trained using the initial training subset.

(2) Twelve radiologists annotated the Retraining1 subset in each CAD software (Table 3). They were split into two groups: board-certified radiologists with more than 5 years of experience and resident radiologists with less experience. The annotation was performed using the CIRCUS DB (Fig. 5). Each lesion was defined as a spherical ROI circumscribing the entire lesion in 3D. Initially, annotation was performed without referencing the results from the CAD software. Subsequently, the annotations were revised by referring to the lesion candidates indicated by the Initial-CAD, which were displayed using spherical ROIs. In the annotation for cerebral aneurysm detection, pre-rendered whole-brain volume-rendered images (Fig. 3c) were also observed using the XTREK VIEW software (J-MAC system, Inc., Sapporo, Japan).

(3) The CAD software retrained by adding the Retraining1 subset annotated by each radiologist is defined as "Retrained-CAD1," resulting in 12 variations for each CAD software.

Table 3 Number of radiologists who annotated Retraining1 and Retraining2 subsets for each CAD software

CAD	Board-certified radiologists	Resident radiologists
Lung nodule detection	5[a]	7[b]
Cerebral aneurysm detection	4	8[c]

[a] Including T.Y., M.K
[b] Including S. Koshino, C.S., M.T., M.N
[c] Including S. Koshino, Y.T., S. Kano

(4) The Retraining2 subset was annotated similarly to Step 2), referencing Retrained-CAD1, which was trained using annotated data from each of the radiologists.
(5) The CAD software retrained by adding the Retraining2 subset annotated by each radiologist is defined as "Retrained-CAD2," resulting in 12 variations for each CAD software.

The performance of each retrained CAD software was evaluated on the test subset. The model of each CAD software was implemented using Python 3.8.5 and PyTorch 1.8.0 [24]. Each model was trained on an NVIDIA DGX A100 server equipped with two AMD Rome 7742 processors (AMD Inc., Santa Clara, CA, USA), 2 TB of memory, and eight graphics processing units (GPUs) (A100 with 40 GB of memory, NVIDIA, Santa Clara, CA, USA). A single GPU was used to train the model.

Integration of annotations from multiple radiologists

We also investigated the performance changes when retraining was performed using integrated annotations from multiple radiologists. For the Retraining1 subset, we integrated annotations as follows: using the product set of annotations from the two radiologists (AND), the sum set of annotations from the two radiologists (OR), and majority voting from the three radiologists (VOTING). VOTING uses annotations integrated as follows:

a. integration by OR between the first and second radiologists
b. integration by AND between the result of a) and the third radiologist

Figure 6 shows examples of ways to integrate annotations among multiple radiologists. We measured the distances between the centroids of spherical ROIs annotated by each radiologist exhaustively. If the distance between the centroids was within 3 mm, the annotations were integrated as the same, and the size and position of the centroid after integration were averaged. The measurements and integration were conducted automatically. The annotating radiologists selected were as follows: 12 radiologists (all radiologists), board-certified radiologists, and resident radiologists. Each CAD software retrained using integrated annotations was evaluated on the test subset.

The free-response operating characteristic (FROC) curve [25, 26], in which sensitivity is plotted against the number of FPs per case (FPs/case), is widely used to evaluate the performance of CAD software. To facilitate comparison between the FROC curves from different types of CAD software in a single number, the competition performance metric (CPM) [27], which defines the average sensitivity at predefined FPs/case (1/8, 1/4, 1/2, 1, 2, 4, and 8 FPs/case) along an FROC curve, was employed as the evaluation criterion. In the comparison between the integration strategy and CAD software, statistical analysis was conducted using the Steel–Dwass test, and a p value of less than 0.05 was considered statistically significant. Statistical analyses were performed using JMP Pro version 17.2.0 (JMP Statistical Discovery LLC, Cary, NC, USA).

Results

Figures 7 and 8 show the performance changes after retraining using the annotated data from each radiologist for lung nodule and cerebral aneurysm detection, respectively. The changes in performance after retraining varied depending on the annotating radiologist. In numerous cases, the CPM increased with the amount of training data. In lung nodule detection, the performance of Retrained-CAD1 or Retrained-CAD2 was sometimes degraded compared to that of the Initial-CAD.

Figure 9 shows the performance of integrating annotations by multiple radiologists for lung nodule detection. The CPMs for retraining using integrated annotations from multiple radiologists tended to be higher than those for retraining using annotations from a single radiologist regardless of the group of annotators. However, there was no significant difference in the Steel–Dwass test results. Figure 10 shows the performance of integrating the annotations by multiple radiologists for cerebral aneurysm detection. The CPMs for retraining using integrated annotations from multiple radiologists were lower than those for retraining using annotations from a single radiologist regardless of the group of annotators. Moreover, in the groups of all radiologists and resident radiologists, significant differences were observed in the Steel–Dwass test results between multiple radiologists and single radiologist.

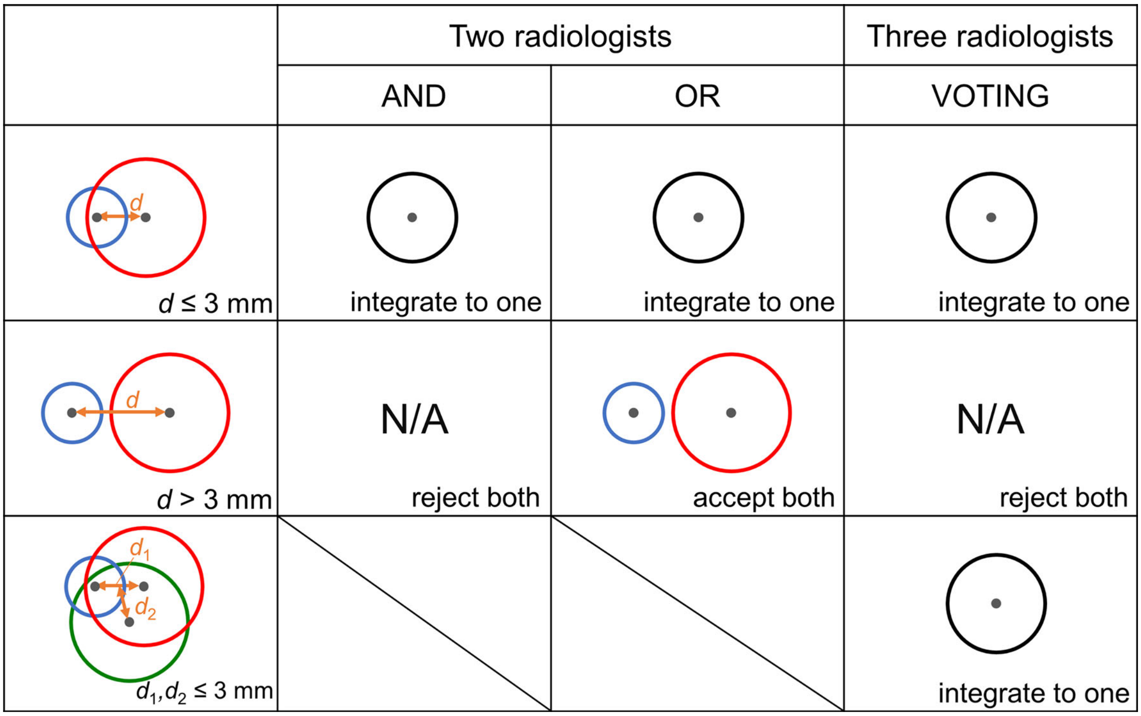


Fig. 6 Examples of ways to integrate annotations among multiple radiologists (The color of circles differs among radiologists. Gray point: centroid of spherical ROI). In integrating annotations (black circle), the size and position of the center of gravity after integration were their averages

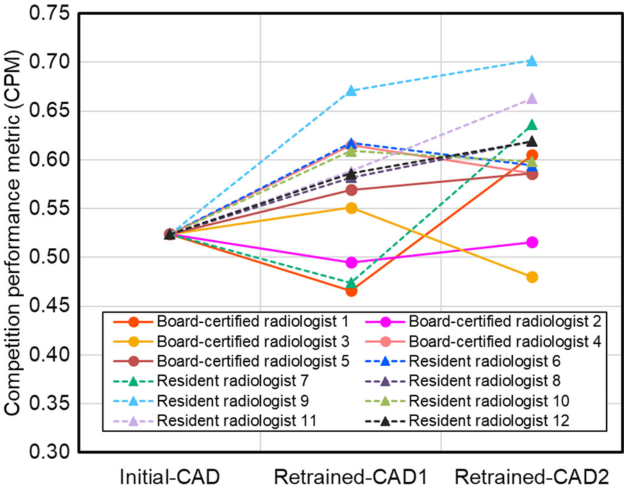


Fig. 7 Performance change after retraining using the annotated data from each radiologist in the lung nodule detection

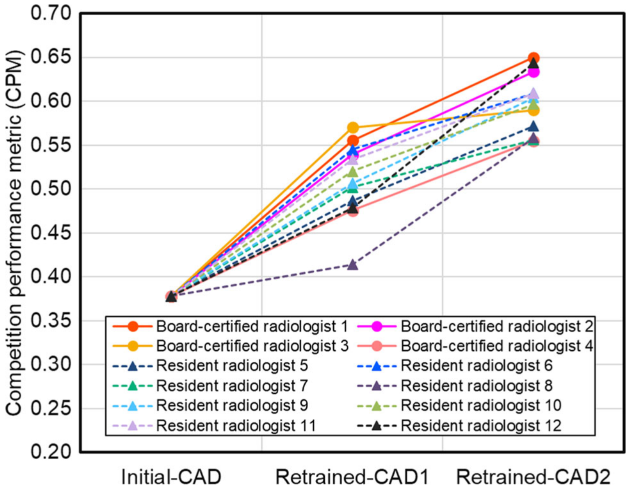


Fig. 8 Performance change after retraining using the annotated data from each radiologist in the cerebral aneurysm detection

Discussion

We experimentally showed that the performance of the CAD software after retraining varied among the annotating radiologists. In addition, for some radiologists, who were Board-certified radiologists 1, 2, and 3 and Resident radiologist 7 in lung nodule detection (Fig. 7), the performance of the CAD software after retraining degraded compared with that of Initial-CAD. This can be attributed to personal tendencies in image diagnosis by annotating radiologists. However, as this was observed by both board-certified radiologists and resident radiologists, it cannot be directly related to the radiologists' years of experience in image interpretation.

The factor contributing to the performance difference of Retraining-CAD1 among annotators (Figs. 7 and 8) is related to the well-known problem of label noise [28, 29] including uncertainty and inconsistency, because the annotation

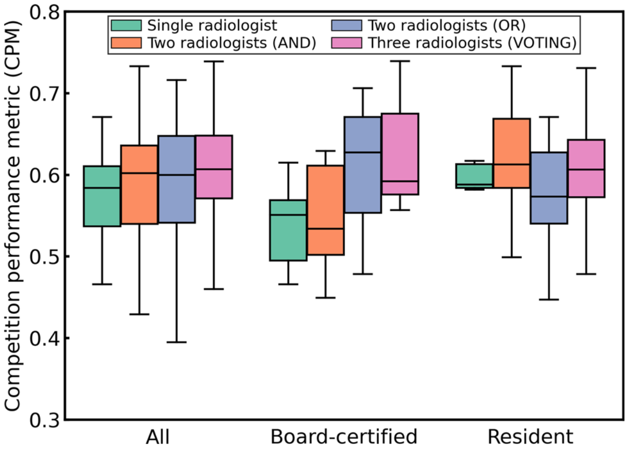


Fig. 9 Performance of integrating annotations by multiple radiologists for lung nodule detection. There was no significant difference in the Steel–Dwass rank sum test result

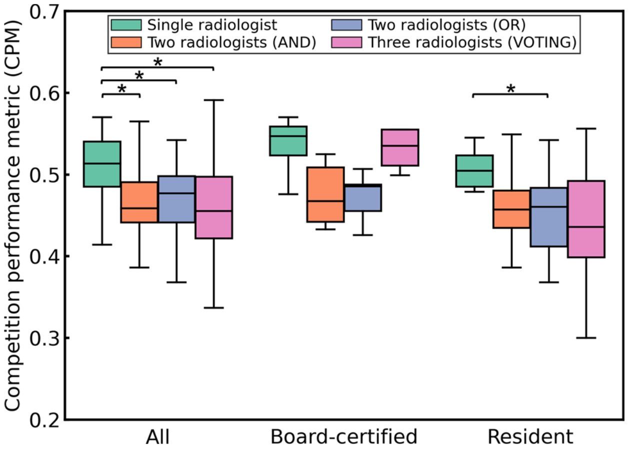


Fig. 10 Performance of integrating annotations by multiple radiologists for cerebral aneurysm detection. *Indicates a p value of less than 0.05

environment for the Retraining1 subset, including the referenced CAD software, is identical. Frénay and Verleysen [28] summarized the taxonomy of label noise in detail. Training machine learning models with label noise are problematic because they can easily overfit corrupted labels, resulting in a lack of generalizability when evaluated on a separate test set. Several research groups have reported methods for handling noisy labels in medical image analyses [30–33]. Xue et al. [30] introduced a global and local representation-guided co-training strategy without refining or relabeling noisy labeled data. Khanal et al. [32] examined the effectiveness of using a self-supervised pretraining approach to improve robustness against noisy labels in a medical image classification task. Penso et al. [33] proposed a calibration procedure for a classification model based on the fact that the confusion matrix of noisy labels can be expressed as the matrix product of the confusion matrix of clean labels and label noise. The performance of retrained CAD software can be improved by applying these methods.

Retraining using annotations integrated from multiple radiologists resulted in different trends in performance depending on the target CAD software (Figs. 9 and 10). Notably, for cerebral aneurysm detection, the performance after retraining using integrated annotations was inferior to that obtained using annotations from a single radiologist. This is not only due to variability in annotations among different radiologists but also depends on the detection sensitivity of the radiologists to the target lesions. According to a preliminary study by our group, the detection sensitivities of the radiologists were 77.4% for pulmonary nodules (5 mm or more in diameter) [34] and 64% for cerebral aneurysms [35]. Although there is no significant difference in performance among the ways of integrating annotations across any target CAD software or group of annotators, in the groups of board-certified radiologists, VOTING showed a tendency to be superior among the three ways for integrating annotations (Figs. 8 and 9). Among the three ways of integrating annotations, AND and OR are simple integrations of the two annotators. By contrast, VOTING is similar to the annotation procedure for the initial training and test subsets, in which a third radiologist resolves any discrepancies after the first two have made their annotations. Abdalla et al. stated that although majority voting is the most commonly used annotation method, it may not be the most suitable [36]. For instance, in the case of screening software, it may be worthwhile considering using a labeling method that maximizes sensitivity, such as OR. If agreement among annotators is desired, adjudicative labeling methods [37] can improve agreement. In addition, Abdalla et al. noted that when the "hard" labeling method by majority voting generates false certainty or noise, it should be expressed using "soft" labels instead [36]. Consequently, integrating annotations from multiple radiologists does not necessarily enhance the quality of the annotations and remains an open problem. We plan to apply soft labels for retraining using annotations integrated from multiple radiologists.

In the development of CAD software using machine learning, it is desirable to improve the performance by repeating the retraining at appropriate intervals. However, as shown by the results of this study, there is a possibility that performance may degrade after retraining. The management of CAD software after retraining remains an open problem, that is, how to monitor performance changes after retraining, and when and how to intervene when performance decline is suspected. Establishing well-defined quality assurance procedures is necessary to monitor the performance of CAD software through retraining. Furthermore, when continuous learning [38] is applied, humans cannot monitor continuously. Therefore, semi-automated or fully automated tools are essential for monitoring the quality and consistency of CAD software after retraining.

Our study had several limitations. First, the lesions were annotated as spherical ROIs to reduce the burden on the annotating radiologists. The results may differ if different types of annotations are used, such as contour drawings and pixel-by-pixel paintings. Second, the annotations from multiple radiologists were automatically integrated. The results may differ from the integration by the consensus of radiologists. Third, annotations for the initial training and test subsets of each CAD software were conducted by three board-certified radiologists. In this case, the uncertainty of disagreement [31], a common type of label noise in medical images, has become a problem. This uncertainty is evident from the results shown in Figs. 7 and 8. To reduce this uncertainty, Drukker et al. stated that testing model performance should not only focus on testing against a variety of independent datasets but also, if possible, against an independent pool of annotators [39]. Tachibana et al. proposed that the decision of each annotator is estimated using the machine learning model and a "virtual conference" to achieve consensus from those results [17]. This method could be used as a potential solution to this problem.

Conclusion

We investigated the changes in performance in the retraining of the CAD software by adding cases stratified by the years of experience of the radiologists who performed the annotation. From our results, we found no direct correlation between the performance and years of experience of the radiologists, although the performance of the CAD software after retraining varied among the annotating radiologists. In addition, retraining using annotations integrated from multiple radiologists may lead to decreased performance compared to that from a single radiologist.

Acknowledgements The Department of Computational Diagnostic Radiology and Preventive Medicine, The University of Tokyo Hospital, is sponsored by HIMEDIC Inc. and Siemens Healthcare K.K. This study was supported by the Japan Agency for Medical Research and Development (AMED) under Grant Numbers 20mk0102148h0202 and 21mk0102148h0203.

Funding Open Access funding provided by The University of Tokyo.

Declarations

Conflict of interest The authors declare no conflicts of interest with regard to the present study.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1975 Declaration of Helsinki, as revised in 2008(5). For this type of study, formal consent is not required. The study was approved by the Research Ethics Committee of the Faculty of Medicine of the University of Tokyo (serial number: 2019274NI-(7), approval date: 6 May 2020, last renewal date: 10 September 2023).

References

1. Giger ML, Chan HP, Boone J (2008) Anniversary paper: history and status of CAD and quantitative image analysis: the role of medical physics and AAPM. Med Phys 35(12):5799–5820
2. van Ginneken B, Schaefer-Prokop CM, Prokop M (2011) Computer-aided diagnosis: how to move from the laboratory to the clinic. Radiology 261(3):719–732
3. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88
4. Yasaka K, Akai H, Kunimatsu A, Kiryu S, Abe O (2018) Deep learning with convolutional neural network in radiology. Jpn J Radiol 36(4):257–272
5. Fujita H (2020) AI-based computer-aided diagnosis (AI-CAD): the latest review to read first. Radiol Phys Technol 13(1):6–19
6. Nomura Y, Miki S, Hayashi N, Hanaoka S, Sato I, Yoshikawa T, Masutani Y, Abe O (2020) Novel platform for development, training, and validation of computer-assisted detection/diagnosis software. Int J Comput Assist Radiol Surg 15(4):661–672
7. Summers RM, Handwerker LR, Pickhardt PJ, Van Uitert RL, Deshpande KK, Yeshwant S, Yao J, Franaszek M (2008) Performance of a previously validated CT colonography computer-aided detection system in a new patient population. AJR Am J Roentgenol 191(1):168–174
8. Gruszauskas NP, Drukker K, Giger ML, Chang RF, Sennett CA, Moon WK, Pesce LL (2009) Breast US computer-aided diagnosis system: robustness across urban populations in South Korea and the United States. Radiology 253(3):661–671
9. Nomura Y, Masutani Y, Hayashi N, Miki S, Nemoto M, Hanaoka S, Yoshikawa T, Ohtomo K (2012) Additional learning of CAD software based on multicenter trial in teleradiology environment. Int J Comput Assist Radiol Surg 7(suppl.1):S270–S271
10. Nomura Y, Masutani Y, Miki S, Hanaoka S, Nemoto M, Yoshikawa T, Hayashi N, Ohtomo K Training strategy for performance improvement in computer-assisted detection of lesions: based on multi-institutional study in teleradiology environment. In: First international symposium on computing and networking (CANDAR 2013), pp 320–323
11. Gibson E, Hu Y, Ghavami N, Ahmed HU, Moore C, Emberton M, Huisman HJ, Barratt DC (2018) Inter-site variability in prostate segmentation accuracy using deep learning. In: MICCAI 2018, LNCS vol 11073, pp 506–514

12. Nomura Y, Hanaoka S, Nakao T, Hayashi N, Yoshikawa T, Miki S, Watadani T, Abe O (2021) Performance changes due to differences in training data for cerebral aneurysm detection in head MR angiography images. Jpn J Radiol 39(11):1039–1048
13. Guan H, Liu Y, Yang E, Yap PT, Shen D, Liu M (2021) Multi-site MRI harmonization via attention-guided deep domain adaptation for brain disorder identification. Med Image Anal 71:102076
14. Shimada K, Daisaki H, Higashiyama S, Kawabe J, Nakaoka R, Shimizu A (2023) Simulation of postmarket fine-tuning of a computer-aided detection system for bone scintigrams and its performance analysis. Adv Biomed Eng 12:51–63
15. Armato SG 3rd, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, Zhao B, Aberle DR, Henschke CI, Hoffman EA, Kazerooni EA, MacMahon H, van Beeke EJ, Yankelevitz D, Biancardi AM, Bland PH, Brown MS, Engelmann RM, Laderach GE, Max D, Pais RC, Qing DP, Roberts RY, Smith AR, Starkey A, Batrah P, Caligiuri P, Farooqi A, Gladish GW, Jude CM, Munden RF, Petkovska I, Quint LE, Schwartz LH, Sundaram B, Dodd LE, Fenimore C, Gur D, Petrick N, Freymann J, Kirby J, Hughes B, Casteele AV, Gupte S, Sallamm M, Heath MD, Kuhn MH, Dharaiya E, Burns R, Fryd DS, Salganicoff M, Anand V, Shreter U, Vastagh S, Croft BY (2011) The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys 38(2):915–931
16. Tan J, Pu J, Zheng B, Wang X, Leader JK (2010) Computerized comprehensive data analysis of lung imaging database consortium (LIDC). Med Phys 37(7):3802–3808
17. Tachibana Y, Nishimori M, Kitamura N, Umehara K, Ota J, Obata T, Higashi T (2020) A neural network model that learns differences in diagnosis strategies among radiologists has an improved area under the curve for aneurysm status classification in magnetic resonance angiography image series. arXiv:2002.01891
18. Nomura Y, Hayashi N, Hanaoka S, Takenaga T, Nemoto M, Miki S, Yoshikawa T, Abe O (2019) Can the spherical gold standards be used as an alternative to painted gold standards for the computerized detection of lesions using voxel-based classification? Jpn J Radiol 37(3):264–273
19. Nomura Y, Nemoto M, Masutani Y, Hanaoka S, Yoshikawa T, Miki S, Maeda E, Hayashi N, Yoshioka N, Ohtomo K (2014) Reduction of false positives at vessel bifurcations in computerized detection of lung nodules. J Biomed Graph Comput 4(3):36–46
20. Hara K, Kataoka H, Satoh Y (2017) Learning spatio-temporal features with 3D residual networks for action recognition. In: 2017 IEEE international conference on computer vision workshops (ICCVW), pp 3154–3160
21. Nyul LG, Udupa JK (1999) On standardizing the MR image intensity scale. Magn Reson Med 42(6):1072–1081
22. Hanaoka S, Nomura Y, Nemoto M, Miki S, Yoshikawa T, Hayashi N, Ohtomo K, Masutani Y, Shimizu A (2015) HoTPiG: a novel geometrical feature for vessel morphometry and its application to cerebral aneurysm detection. In: MICCAI 2015, LNCS vol 9350, pp 103–110
23. Nakao T, Hanaoka S, Nomura Y, Sato I, Nemoto M, Miki S, Maeda E, Yoshikawa T, Hayashi N, Abe O (2018) Deep neural network-based computer-assisted detection of cerebral aneurysms in MR angiography. J Magn Reson Imaging 47(4):948–953
24. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (2019) PyTorch: an imperative style, high-performance deep learning library. Adv Neural Inf Process Syst 32:8024–8035
25. Chakraborty DP, Berbaum KS (2004) Observer studies involving detection and localization: modeling, analysis, and validation. Med Phys 31(8):2313–2330
26. Metz CE (2006) Receiver operating characteristic analysis: a tool for the quantitative evaluation of observer performance and imaging systems. J Am Coll Radiol 3(6):413–422
27. Setio AAA, Traverso A, de Bel T, Berens MSN, Bogaard CVD, Cerello P, Chen H, Dou Q, Fantacci ME, Geurts B, Gugten RV, Heng PA, Jansen B, de Kaste MMJ, Kotov V, Lin JY, Manders J, Sonora-Mengana A, Garcia-Naranjo JC, Papavasileiou E, Prokop M, Saletta M, Schaefer-Prokop CM, Scholten ET, Scholten L, Snoeren MM, Torres EL, Vandemeulebroucke J, Walasek N, Zuidhof GCA, Ginneken BV, Jacobs C (2017) Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Med Image Anal 42:1–13
28. Frénay B, Verleysen M (2014) Classification in the presence of label noise: a survey. IEEE Trans Neural Netw Learn Syst 25(5):845–869
29. Song H, Kim M, Park D, Shin Y, Lee JG (2023) Learning from noisy labels with deep neural networks: a survey. IEEE Trans Neural Netw Learn Syst 34(11):8135–8153
30. Xue C, Yu L, Chen P, Dou Q, Heng PA (2022) Robust medical image classification from noisy labeled data with global and local representation guided co-training. IEEE Trans Med Imaging 41(6):1371–1382
31. Ju L, Wang X, Wang L, Mahapatra D, Zhao X, Zhou Q, Liu T, Ge Z (2022) Improving medical images classification with label noise using dual-uncertainty estimation. IEEE Trans Med Imaging 41(6):1533–1546
32. Khanal B, Bhattarai B, Khanal B, Linte CA (2023) Improving medical image classification in noisy labels using only self-supervised pretraining. DEMI 2023, LNCS 14314:78–90
33. Penso C, Frenkel L, Goldberger J (2024) Confidence calibration of a medical imaging classification system that is robust to label noise. IEEE Trans Med Imaging. https://doi.org/10.1109/TMI.2024.3353762
34. Miki S, Nomura Y, Hayashi N, Hanaoka S, Maeda E, Yoshikawa T, Masutani Y, Abe O (2021) Prospective study of spatial distribution of missed lung nodules by readers in CT lung screening using computer-assisted detection. Acad Radiol 28(5):647–654
35. Miki S, Hayashi N, Masutani Y, Nomura Y, Yoshikawa T, Hanaoka S, Nemoto M, Ohtomo K (2016) Computer-assisted detection of cerebral aneurysms in MR angiography in a routine image-reading environment: effects on diagnosis by radiologists. AJNR Am J Neuroradiol 37(6):1038–1043
36. Abdalla M, Fine B (2023) Hurdles to artificial intelligence deployment: noise in schemas and “gold” labels. Radiol Artif Intell 5(2):e220056
37. Duggan GE, Reicher JJ, Liu Y, Tse D, Shetty S (2021) Improving reference standards for validation of AI-based radiography. Br J Radiol 94(1123):20210435
38. Pianykh OS, Langs G, Dewey M, Enzmann DR, Herold CJ, Schoenberg SO, Brink JA (2020) Continuous learning AI in radiology: implementation principles and early applications. Radiology 297(1):6–14
39. Drukker K, Chen W, Gichoya J, Gruszauskas N, Kalpathy-Cramer J, Koyejo S, Myers K, Sá RC, Sahiner B, Whitney H, Zhang Z, Giger M (2023) Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. J Med Imaging 10(6):061104