Original Article | Artificial Intelligence

eISSN 2005-8330
https://doi.org/10.3348/kjr.2023.1027
Korean J Radiol 2023;24(12):1179-1189

Korean Journal of Radiology
KJR

Reporting Quality of Research Studies on AI Applications in Medical Images According to the CLAIM Guidelines in a Radiology Journal With a Strong Prominence in Asia

Dong Yeong Kim[1]*, Hyun Woo Oh[2]*, Chong Hyun Suh[1]

[1]Department of Radiology and Research Institute of Radiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
[2]NAVER Inc., Seongnam, Republic of Korea

Objective: We aimed to evaluate the reporting quality of research articles that applied deep learning to medical imaging. Using the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) guidelines and a journal with prominence in Asia as a sample, we intended to provide an insight into reporting quality in the Asian region and establish a journal-specific audit.
Materials and Methods: A total of 38 articles published in the *Korean Journal of Radiology* between June 2018 and January 2023 were analyzed. The analysis included calculating the percentage of studies that adhered to each CLAIM item and identifying items that were met by ≤ 50% of the studies. The article review was initially conducted independently by two reviewers, and the consensus results were used for the final analysis. We also compared adherence rates to CLAIM before and after December 2020.
Results: Of the 42 items in the CLAIM guidelines, 12 items (29%) were satisfied by ≤ 50% of the included articles. None of the studies reported handling missing data (item #13). Only one study respectively presented the use of de-identification methods (#12), intended sample size (#19), robustness or sensitivity analysis (#30), and full study protocol (#41). Of the studies, 35% reported the selection of data subsets (#10), 40% reported registration information (#40), and 50% measured inter and intrarater variability (#18). No significant changes were observed in the rates of adherence to these 12 items before and after December 2020.
Conclusion: The reporting quality of artificial intelligence studies according to CLAIM guidelines, in our study sample, showed room for improvement. We recommend that the authors and reviewers have a solid understanding of the relevant reporting guidelines and ensure that the essential elements are adequately reported when writing and reviewing the manuscripts for publication.
Keywords: Reporting quality; Artificial intelligence; Medical imaging; CLAIM guidelines; Asia

INTRODUCTION

With an increasing number of artificial intelligence (AI) publications in the field of medical imaging, it becomes imperative to have evidence-based guidelines to unify the reporting of AI studies [1,2]. Owing to their complex methodology and weak reporting qualities, AI studies are perceived as challenging for readers [3-5]. To address this issue, initial protocols were derived from the standards set for randomized clinical trials. Specifically, the Consolidated Standards of Reporting Trials-AI (CONSORT-AI), Standard Protocol Items: Recommendations for Interventional Trials-AI (SPIRIT-AI) guidelines, and Developmental and Exploratory Clinical Investigations of DEcision support systems driven by AI (DECIDE-AI) were originally developed for consistent reporting of clinical trials and their protocols [6-8]. In addition, Standards for Reporting of Diagnostic

Received: October 20, 2023 **Revised:** October 25, 2023
Accepted: October 26, 2023
*These authors contributed equally to this work.
Corresponding author: Chong Hyun Suh, MD, PhD, Department of Radiology and Research Institute of Radiology, Asan Medical Center, University of Ulsan College of Medicine, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Republic of Korea
• E-mail: chonghyunsuh@amc.seoul.kr

Accuracy Study-AI (STARD-AI) and Transparent Reporting of a multivariable prediction model of Individual Prognosis Or Diagnosis-AI (TRIPOD-AI) are currently under development [9,10].

Simultaneously, the Checklist for Artificial Intelligence in Medical Imaging (CLAIM), a comprehensive guideline covering the broad application of AI in medical imaging with an emphasis on model development, was published [11] (Table 1). Based on these advantages, CLAIM is one of the best practice guidelines for AI-supported medical imaging research [12] and is endorsed by the Radiological Society of North America (RSNA) journals, which are one of the premier journal groups in medical imaging.

To the best of our knowledge, assessments of the quality of reporting across a broad spectrum of AI medical imaging studies using the CLAIM guidelines as an evaluation tool are limited. Previous studies addressing adherence to CLAIM in AI studies do not fully utilize the advantages of the CLAIM guidelines because they limit the topic to specific disease entities [13-18] or specific AI application methods [19,20]. In addition, although several AI reporting guidelines, including CLAIM, are primarily developed in Western countries, AI research is also being actively conducted in Asia, most notably by Chinese and Korean researchers [21]. Therefore, analyzing studies with various applications and disease entities from countries distant from the guidelines' epicenters could be uniquely informative, for which the *Korean Journal of Radiology* (KJR) could be a good sample, as it is a broad-spectrum general radiology journal with a reputation and strong presence in contributions from Asia. Moreover, such an analysis would provide a valuable journal-specific audit.

This study aimed to evaluate the reporting quality of research articles that have applied deep learning to medical imaging. For this assessment, we used the CLAIM guidelines. We selected a journal with prominence in Asia as a sample, with the intention of providing insight into reporting quality in the Asian region and, also establishing a journal-specific audit.

MATERIALS AND METHODS

Overview of CLAIM

CLAIM is based on the STARD guidelines and has been expanded to cover AI applications in medical imaging, such as classification, image reconstruction, text analysis, and workflow optimization. It focuses on AI model development and particularly emphasizes the generalizability of research [11].

Literature Search Strategy and Study Selection

Using the MEDLINE database, we searched for all potential articles discussing AI applications in medical imaging published in a single peer-reviewed journal, the KJR, between June 2018 and January 2023. The search terms were ((“artificial intelligence”) OR (“deep learning”) OR ("machine learning") OR (“convolutional neural network”) OR (“deep neural network”)) AND (“Korean Journal of Radiology” [Journal]). The search date was May 10, 2023. A total of 83 records (i.e., abstracts and titles) were identified from the MEDLINE database, and two reviewers (D.Y.K., with 2 years of experience in radiology, and H.W.O., with 4 years of experience in AI development and research) evaluated the eligibility of each article. Among them the following papers were excluded: 17 review articles, 10 editorials, 5 non-AI studies, and 1 paper each of survey, case report, or erratum.

After the first screening, the eligibility of the remaining 48 studies was evaluated. Four records were excluded because they did not use deep learning. Six records were excluded because they were not related to model development or validation studies. As a result, full texts from 38 studies were included in the analysis [22-59] (Fig. 1).

Data Extraction

The following data were independently extracted from the included articles: name of the first author, year of publication, type of AI application (classification, detection, segmentation, or image reconstruction), and study objective (model development or validation). In addition, for each article, we evaluated the sections (title, abstract, introduction, methods, results, discussion, and other information) of the included papers according to CLAIM and referred to the detailed topics (such as study design, data, ground truth, data partitions, and model) of CLAIM to evaluate the methods and results sections. Items #9 to #13 and #20 to #27, which belong to the methods section, were evaluated only in articles on model development. Data were independently extracted by two reviewers (D.Y.K. and H.W.O.). If a disagreement occurred, a third reviewer (C.H.S., with 10 years of experience in performing systematic reviews) was consulted to reach a consensus.

Data Analysis

We identified the CLAIM checklist items to which ≤ 50% of the articles adhered. We grouped these items and a few

https://doi.org/10.3348/kjr.2023.1027

Korean Journal of Radiology
KJR

Table 1. Adherence to CLAIM checklist

Section and Topic	Item #	Checklist item	Number of articles adhered
Title or Abstract			
	1	Identification as a study of AI methodology, specifying the category of technology used (e.g., deep learning)	36/38 (95%)
	2	Structured summary of study design, methods, results, and conclusions	38/38 (100%)
Introduction			
	3	Scientific and clinical background, including the intended use and clinical role of the AI approach	38/38 (100%)
	4	Study objectives and hypotheses	38/38 (100%)
Methods			
Study design	5	Prospective or retrospective study	36/38 (95%)
	6	Study goal, such as model creation, exploratory study, feasibility study, non-inferiority trial	38/38 (100%)
Data	7	Data sources	38/38 (100%)
	8	Eligibility criteria: how, where, and when potentially eligible participants or studies were identified (e.g., symptoms, results from previous tests, inclusion in registry, patient-care setting, location, dates)	34/38 (89%)
	9	Data pre-processing steps	18/23 (78%)
	10	Selection of data subsets, if applicable	8/23 (35%)
	11	Definitions of data elements, with references to Common Data Elements	NA
	12	De-identification methods	1/23 (4%)
	13	How missing data were handled	0/23 (0%)
Ground truth	14	Definition of ground truth reference standard, in sufficient detail to allow replication	26/27 (96%)
	15	Rationale for choosing the reference standard (if alternatives exist)	7/10 (70%)
	16	Source of ground-truth annotations; qualifications and preparation of annotators	15/20 (75%)
	17	Annotation tools	14/18 (78%)
	18	Measurement of inter- and intrarater variability; methods to mitigate variability and/or resolve discrepancies	9/18 (50%)
Data partitions	19	Intended sample size and how it was determined	1/38 (3%)
	20	How data were assigned to partitions; specify proportions	21/23 (91%)
	21	Level at which partitions are disjoint (e.g., image, study, patient, institution)	21/23 (91%)
Model	22	Detailed description of model, including inputs, outputs, all intermediate layers and connections	20/23 (87%)
	23	Software libraries, frameworks, and packages	11/23 (48%)
	24	Initialization of model parameters (e.g., randomization, transfer learning)	9/23 (39%)
Training	25	Details of training approach, including data augmentation, hyperparameters, number of models trained	14/23 (61%)
	26	Method of selecting the final model	5/23 (22%)
	27	Ensembling techniques, if applicable	0/23 (0%)
Evaluation	28	Metrics of model performance	38/38 (100%)
	29	Statistical measures of significance and uncertainty (e.g., confidence intervals)	37/38 (97%)
	30	Robustness or sensitivity analysis	1/38 (3%)
	31	Methods for explainability or interpretability (e.g., saliency maps), and how they were validated	26/30 (87%)
	32	Validation or testing on external data	27/34 (79%)
Results			
Data	33	Flow of participants or cases, using a diagram to indicate inclusion and exclusion	31/38 (82%)
	34	Demographic and clinical characteristics of cases in each partition	26/38 (68%)
Model performance	35	Performance metrics for optimal model(s) on all data partitions	38/38 (100%)
	36	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	37/38 (97%)

https://doi.org/10.3348/kjr.2023.1027

Table 1. Adherence to CLAIM checklist (continued)

Section and Topic	Item #	Checklist item	Number of articles adhered
	37	Failure analysis of incorrectly classified cases	16/22 (73%)
Discussion			
	38	Study limitations, including potential bias, statistical uncertainty, and generalizability	38/38 (100%)
	39	Implications for practice, including the intended use and/or clinical role	38/38 (100%)
Other information			
	40	Registration number and name of registry	2/5 (40%)
	41	Where the full study protocol can be accessed	1/38 (3%)
	42	Sources of funding and other support; role of funders	34/38 (89%)

CLAIM = Checklist for Artificial Intelligence in Medical Imaging, AI = artificial intelligence, NA = not applicable

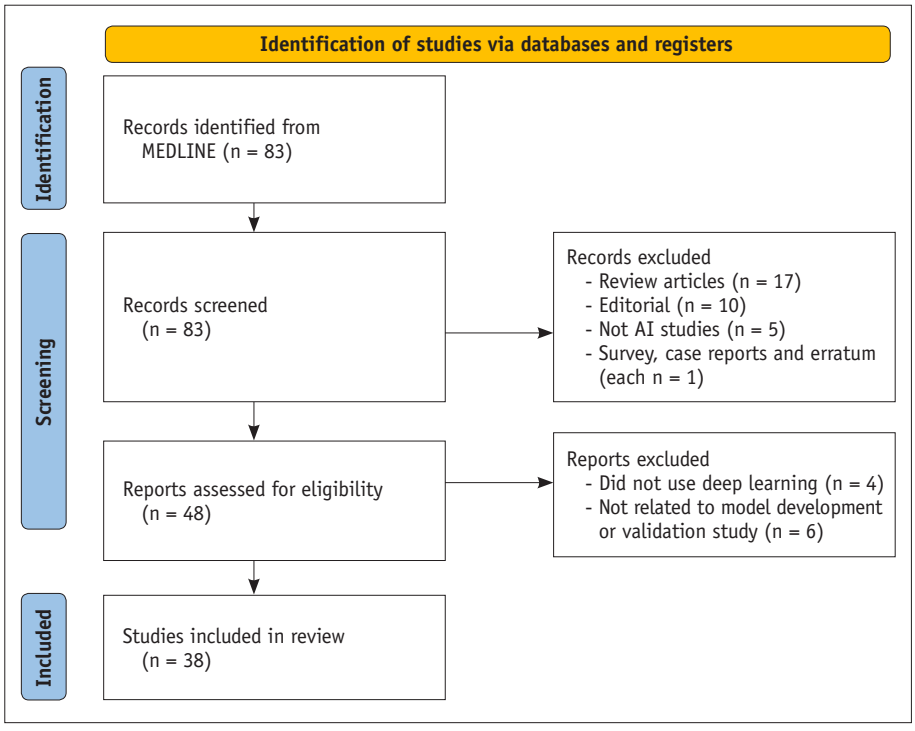

Fig. 1. Flow diagram of the study selection process. AI = artificial intelligence

additional items (unless included in the "≤ 50% adherence rate" items) we wanted to review further, into 5 relevant domains (Table 2). Suggestions for enhancing the quality of medical imaging articles involving AI were provided based on these established domains. We also compared the rates of adherence to the CLAIM checklist (published in March 2020) between articles published up to December 2020 and those published in January 2021 and later using the chi-squared test.

RESULTS

Characteristics of the Included Studies

The characteristics of the 38 included studies are summarized in Supplementary Table 1. In terms of the type of AI application, 9 studies (24%) applied AI for classification [22,27,31,32,38,39,43,55,57], 6 studies (16%) for detection [35,36,49-51,58], 11 studies (29%) for segmentation [25,28,29,33,37,40,41,45,46,48,53], and 12 studies (31%) for image reconstruction [23,24,26,30,34,42,44,47,52,54,56,59]. The study objective of 23 papers (61%) was classified as model development [22,23,25,26,28-32,35-39,42,43,45,46,48,49,53,55,59], and the remaining 15 papers (39%) were classified as validation studies [24,27,33,34,40,41,44,47,50-52,54,56-58]. Five studies (13%) were prospective [22,41,42,55,58] and 33 studies (87%) were retrospective [23-40,43-54,56,57,59].

https://doi.org/10.3348/kjr.2023.1027

Korean Journal of Radiology
KJR

Table 2. Grouping of the items with ≤ 50% adherence rate and additional items (#22 and 25)

Domain	CLAIM items	Checklist
1. Information concerning data and data partitioning	#10, #12, #13, #19	1. Report detailed information about data and data partitions. - If subsets of the raw data were used, report detailed information. - Report the methods of de-identification. - State clearly how missing data were handled. 2. Mention the intended sample size and a reference about how to calculate sample size.
2. Concrete description about ground truth	#18	3. Provide methods for evaluating inter- and intrarater variability and ways to resolve them.
3. Details concerning model and training	#22, #23, #24, #25, #26, #27	4. Provide techniques to construct and train AI model for reproducibility and transparency. - Report a detailed structure of the model and the name of software libraries. - Indicate parameter initialization methods. - Describe all the training procedures and hyperparameters. - State model selection method and ensemble method, if applicable.
4. Evaluating model performance	#30	5. Perform robustness or sensitivity analysis to ensure software to keep an "acceptable" behavior, in spite of exceptional or unforeseen execution conditions.
5. Other information	#40, #41	6. If the current study is a prospective study, it is recommended to provide registration information. 7. Share all computer code used for modeling or data analysis in a publicly accessible repository.

CLAIM = Checklist for Artificial Intelligence in Medical Imaging, AI = artificial intelligence

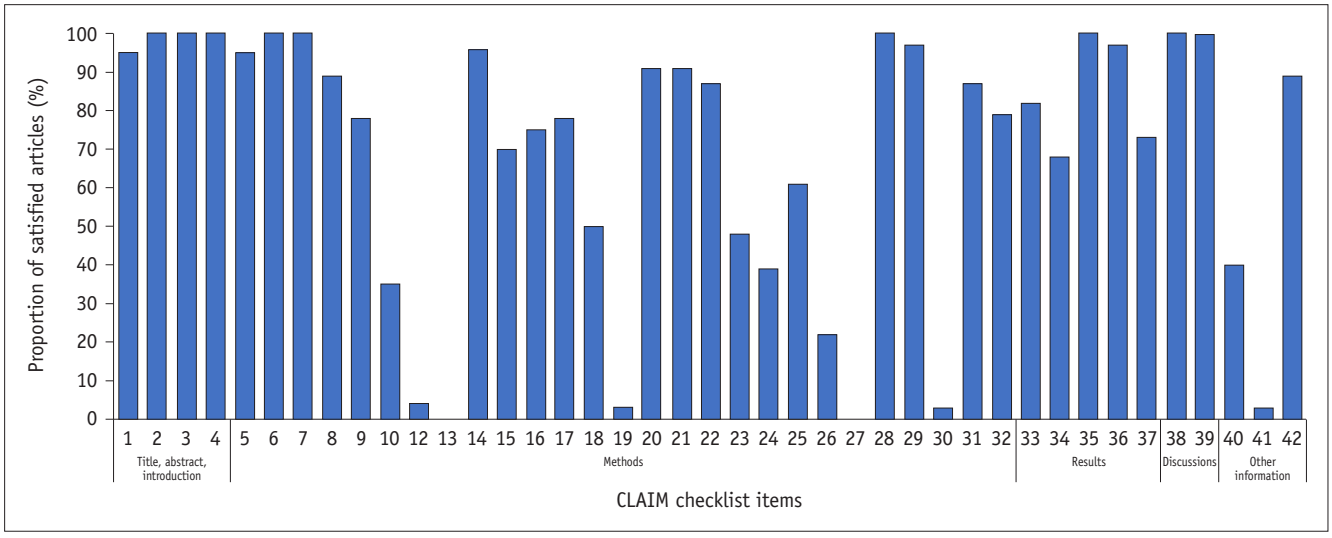

Fig. 2. Bar chart demonstrating the proportion of articles that satisfied each item of the CLAIM checklist. Twelve items (29%) were reported in ≤ 50% of the articles. CLAIM = Checklist for Artificial Intelligence in Medical Imaging

Adherence to CLAIM

The included articles were evaluated for adherence to each item of the CLAIM checklist (Table 1). Of the 42 guideline items, 12 items (29%) were reported in ≤ 50% of the articles (Fig. 2). Most of the studies met the criteria for the title, abstract, introduction, and discussion sections, but there were frequent instances of incomplete reporting in the methods and other information sections, particularly concerning the data topic. None of the studies reported handling missing data (item #13). Only one study respectively reported the use of de-identification methods (item #12), intended sample size (item #19), robustness or sensitivity analysis (item #30), and the full study protocol (item #41). Thirty five percent of the studies reported the selection of data

 https://doi.org/10.3348/kjr.2023.1027

subsets (item #10), 40% reported registration information (item #40), and 50% measured inter and intrarater variability (item #18).

For the 12 items that were met by less than 50% of the included articles, there were no significant changes in the adherence rates between the two periods. There was a slight increase in adherence to data partitions, model, and training-related items (#20, #21, and #22) in the methods section, but without a statistical difference, while item #25, related to detailed reporting on the training procedure, showed a statistically significant improvement (33% to 79%; $P = 0.031$). In the results section, item #34, which was regarding reporting demographic and clinical characteristics for each data partition, also showed marginal improvement (45% to 78%; $P = 0.05$). We grouped the 12 items with ≤ 50% adherence rate and two additional items (#22 and #25) that we wanted to emphasize, into 5 domains, as listed below (Table 2).

Information Concerning Data and Data Partitioning (Items #10, #12, #13, and #19)

The use of data subsets, such as data cropping, focusing on a specific segment of the dataset, could facilitate model training and testing [60]. According to the CLAIM guidelines, the use of data subsets should be indicated when applicable. It was observed that 35% of the studies employed this method in their research [22,26,31,36,39,43,48,55]. De-identification is an important ethical aspect in AI research. In our review, only one study (4%) explicitly reported the use of anonymization [31]. In the reviewed articles, the target tasks were primarily related to computer vision such as image classification, image segmentation, detection, and image recognition. Owing to the nature of these tasks, there was no mention of techniques for handling missing data, which are more commonly associated with tabular data. Item #19 addresses the intended sample size. Only one article (3%) met these criteria [51]. It mentioned the intended sample size and provided reference for calculating the sample size tables for receiver operating characteristic studies.

Concrete Descriptions about Ground Truth (Item #18)

Item #18 concerns measuring inter and intrarater variability and the method to resolve it. This could not be applied to 20 papers. Therefore, item #18 was evaluated for 18 papers altogether, and 50% of these articles met the criteria [32,38,39,44-46,48,49,58]. Two studies [32,39] did not suggest methods to assess inter and intrarater variability, whereas seven studies [38,44-46,48,49,58] did not report methods to reduce or mitigate this variability or resolve discrepancies.

Details Concerning Model and Training (Items #22, #23, #24, #25, #26, #27)

In the quest for reproducibility and transparency in the field of AI research, a comprehensive and detailed description of an AI model's structure is a critical element. CLAIM requests a 'complete detailed structure': the components of input and output, the structure of the neural network including pooling, normalization, regularization, and activation layer. It was found that 20 out of 23 (87%) articles provided a detailed structure of their proposed model [22,23,25,26,28,29,32,35-37,39,42,43,45,46,48,49,53,55,59]. Among the three articles that did not provide enough details, one cited a previous paper for its model structure [30], whereas the other two articles lacked sufficient detail. Software libraries (item #23), initialization of model parameters (item #24), details of the training approach (item #25), method of selecting the final model (item #26), and ensembling techniques (item #27) are described in the Supplementary Material.

Evaluating Model Performance (Item #30)

Item #30 concerns robustness or sensitivity analysis. Among the included articles, only one (3%) [25] satisfied the criteria. The paper mentioned that subgroups of various clinical conditions were included, and several types of computed tomography scanners were used to develop a robust deep-learning algorithm.

Other Information (Items #40 and #41)

Item #40 is related to clinical trial registration, but most of the included studies were retrospective. Therefore, only five studies could be evaluated, and 40% of these studies reported registration information [41,58]. CLAIM emphasizes that authors should share all computer code used for modeling or data analysis in a publicly accessible repository; in this aspect, item #41 was satisfied by only one article (3%) [53]. Another study [56] mentioned that all data generated or analyzed were included in the text and supplements, but the computer code was not publicly disclosed.

https://doi.org/10.3348/kjr.2023.1027

Korean Journal of Radiology
KJR

DISCUSSION

Our study revealed the focus areas for improving the reporting quality of studies on AI applications in medical imaging. Although the results were obtained from a single journal, given the status of the journal (including Q1 status according to the Journal Citation Reports™ and Scimago Journal & Country Rank [61,62]), they may serve as a snapshot of the reporting quality among articles generally regarded as high-quality research studies in the field of radiology. Our evaluations can be segmented into five categories, where detailed guidance and recommendations can be provided: 1) information concerning data and data partitioning, 2) concrete descriptions about ground truth, 3) details concerning model and training, 4) evaluating model performance, and 5) other information.

The results of this study were similar to those of previous studies using CLAIM [13,15,16]. In the case of item #18, dealing with inter and intrareader variability, only 50% of the studies satisfied the criteria, and this was also low in previous studies [13,16]. Human perception remains in the initial stage of image reading; however, a radiologist's proficiency depends on multiple factors. Consequently, the outcomes of an imaging technique frequently hinge on the inherent qualities of the observer. To address this issue, it is advised to involve multiple observers and conduct independent readings to gain a comprehensive understanding of potential variations in the results [63].

In addition, the achievement rate of item #30, robustness, or sensitivity analysis, was very low, and similar results have been reported in previous studies [15,16]. Robustness can be defined as the ability of the software to maintain "acceptable" behavior despite exceptional or unforeseen execution conditions [64]. To some extent, the achievements of deep learning models rely on their ability to generalize and remain stable. Studies have demonstrated that these models can produce different outputs when presented with slight variations in input data. Such response variability to minor changes might indicate algorithmic instability, potentially resulting in misclassification and challenges in generalization [65]. Therefore, it is important to evaluate the robustness and stability of AI models before their clinical implementation, especially in the field of medical imaging, and authors should consider reporting them.

In the case of the ground truth topic in the methods section, the report rate exceeded 50%, except for item #18, which had a different result from that in previous studies that showed a low report rate for this item [13,15,16]. While previous studies collected and evaluated specific disease imaging studies, the current study conducted a wide range of evaluations without distinguishing between AI applications and diseases. As a result, many papers dealing with image reconstruction and segmentation were included. Since this is a research area where alternative reference standards are relatively difficult to find, it is judged that the report rate was high because this study was a little less stringent in defining the ground truth than other studies.

Reproducibility and transparency in deep-learning modeling are crucial factors for enhancing the quality of research, and standardized guidelines are necessary to achieve them. However, in the current guidelines, items #22 and #25, may be perceived as vague or overly strict. For example, if a convolution layer is employed as an intermediate layer, many details, such as stride, padding, dilation rate, and bias should be applied [66]. Describing these details thoroughly may be perceived as being redundant or overly strict. It would be advisable to provide further guidelines that specify the methodological details that should be included rather than demanding a full, exhaustive description. Furthermore, the clarity can be enhanced by reorganizing certain items. For instance, item #10, which is related to the preprocessing of the input, could be included in item #22. Part of item #25 concerning the selection method of the best-performing model, could be transferred to item #26. Finally, item #13, the handling of missing data, pertains only to tabular data [67]. Therefore, a modification that mandates this item only in papers using tabular data can enhance the clarity of the guidelines. As mentioned above, there are many opinions regarding improvements through the reorganization or clarification of some items, and an updated CLAIM that reflects these is being developed [12].

Sharing executable algorithms or data in a publicly accessible repository is currently recommended for publication by most peer-reviewed journals, including *Radiology* [68] and not just by the CLAIM guidelines. However, a previous study using CLAIM [16] reported a low adherence rate to this policy. Code sharing facilitates the evaluation of an AI algorithm using data from the intended healthcare system, which is required to confirm the algorithm's generalizability to the user's environment [2,69]. In addition, code sharing can provide users with a deeper insight into the necessary computing power and logistical factors, such as data transfer and image preprocessing [2].

Our study has several limitations. First, we used a specific

 https://doi.org/10.3348/kjr.2023.1027

journal as the study sample; therefore, our results may have limited generalizability and should be viewed along with other similar published studies. Second, CLAIM comprises a multitude of reporting guidelines for AI research studies and has its own limitations. For example, the direction of reporting the sample size does not distinguish between the training and test datasets. While sample size estimation is critical for testing an algorithm with adequate statistical power, a priori estimation of an adequate training data size is not entirely practical or feasible [2]. CLAIM is currently under revision [12], and our study results would need to be updated when new reporting guidelines emerge. Finally, there is potential inexperience in our analysis owing to the inaugural application of the CLAIM criteria. Unfamiliarity with certain aspects of the criteria may have influenced our evaluation.

In conclusion, the reporting quality of AI studies with respect to CLAIM of AI studies in our study sample, showed room for improvement. We recommend that the authors and reviewers have a solid understanding of the relevant reporting guidelines and ensure that the essential elements are adequately reported when writing and reviewing the manuscripts for publication.

Supplement

The Supplement is available with this article at https://doi.org/10.3348/kjr.2023.1027.

Availability of Data and Material

The datasets generated or analyzed during the study are available from the corresponding author on reasonable request.

Conflicts of Interest

Chong Hyun Suh, the Assistant to the Editor of the *Korean Journal of Radiology*, was not involved in the editorial evaluation or decision to publish this article. All authors have declared no conflicts of interest.

Author Contributions

Conceptualization: all authors. Data curation: Dong Yeong Kim, Hyun Woo Oh. Formal analysis: Dong Yeong Kim, Hyun Woo Oh. Funding acquisition: Chong Hyun Suh. Investigation: Chong Hyun Suh. Methodology: Chong Hyun Suh. Project administration: Chong Hyun Suh. Resources: Chong Hyun Suh. Software: Dong Yeong Kim, Hyun Woo Oh. Supervision: Chong Hyun Suh. Validation: Chong Hyun Suh. Visualization: Dong Yeong Kim, Hyun Woo Oh. Writing—original draft: Dong Yeong Kim, Hyun Woo Oh. Writing—review & editing: all authors.

ORCID IDs

Dong Yeong Kim
https://orcid.org/0000-0002-8548-7377
Hyun Woo Oh
https://orcid.org/0009-0006-7809-2808
Chong Hyun Suh
https://orcid.org/0000-0002-4737-0530

Funding Statement

This work was supported by the National Research Foundation of Korea (NRF- 2021R1C1C1014413).

REFERENCES

1. Park SH, Han K, Jang HY, Park JE, Lee JG, Kim DW, et al. Methods for clinical evaluation of artificial intelligence algorithms for medical diagnosis. *Radiology* 2023;306:20-31
2. Park SH, Sul AR, Ko Y, Jang HY, Lee JG. Radiologist's guide to evaluating publications of clinical research on ai: how we do it. *Radiology* 2023;308:e230288
3. Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digit Med* 2021;4:65
4. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 2019;1:e271-e297
5. 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;20:405-410
6. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Lancet Digit Health* 2020;2:e549-e560
7. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Lancet Digit Health* 2020;2:e537-e548
8. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* 2022;377:e070904
9. Sounderajah V, Ashrafian H, Golub RM, Shetty S, De Fauw J,

Hooft L, et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open* 2021;11:e047709
10. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* 2021;11:e048008
11. 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:e200029
12. Tejani AS, Klontzas ME, Gatti AA, Mongan J, Moy L, Park SH, et al. Updating the checklist for artificial intelligence in medical imaging (CLAIM) for reporting AI research. *Nat Mach Intell* 2023;5:950-951
13. Belue MJ, Harmon SA, Lay NS, Daryanani A, Phelps TE, Choyke PL, et al. The low rate of adherence to checklist for artificial intelligence in medical imaging criteria among published prostate MRI artificial intelligence algorithms. *J Am Coll Radiol* 2023;20:134-145
14. Lans A, Pierik RJB, Bales JR, Fourman MS, Shin D, Kanbier LN, et al. Quality assessment of machine learning models for diagnostic imaging in orthopaedics: a systematic review. *Artif Intell Med* 2022;132:102396
15. Si L, Zhong J, Huo J, Xuan K, Zhuang Z, Hu Y, et al. Deep learning in knee imaging: a systematic review utilizing a checklist for artificial intelligence in medical imaging (CLAIM). *Eur Radiol* 2022;32:1353-1361
16. Bhandari A, Scott L, Weilbach M, Marwah R, Lasocki A. assessment of artificial intelligence (AI) reporting methodology in glioma MRI studies using the checklist for ai in medical imaging (CLAIM). *Neuroradiology* 2023;65:907-913
17. Zhong J, Hu Y, Zhang G, Xing Y, Ding D, Ge X, et al. An updated systematic review of radiomics in osteosarcoma: utilizing CLAIM to adapt the increasing trend of deep learning application in radiomics. *Insights Imaging* 2022;13:138
18. Tsang B, Gupta A, Takahashi MS, Baffi H, Ola T, Doria AS. Applications of artificial intelligence in magnetic resonance imaging of primary pediatric cancers: a scoping review and CLAIM score assessment. *Jpn J Radiol* 2023;41:1127-1147
19. Kouli O, Hassane A, Badran D, Kouli T, Hossain-Ibrahim K, Steele JD. Automated brain tumor identification using magnetic resonance imaging: a systematic review and meta-analysis. *Neurooncol Adv* 2022;4:vdac081
20. Alabed S, Maiter A, Salehi M, Mahmood A, Daniel S, Jenkins S, et al. Quality of reporting in AI cardiac MRI segmentation studies - a systematic review and recommendations for future studies. *Front Cardiovasc Med* 2022;9:956811
21. Guo Y, Hao Z, Zhao S, Gong J, Yang F. Artificial intelligence in health care: bibliometric analysis. *J Med Internet Res* 2020;22:e18228
22. Choi JS, Han BK, Ko ES, Bae JM, Ko EY, Song SH, et al. Effect of a deep learning framework-based computer-aided diagnosis system on the diagnostic performance of radiologists in differentiating between malignant and benign masses on breast ultrasonography. *Korean J Radiol* 2019;20:749-758
23. Lee SM, Lee JG, Lee G, Choe J, Do KH, Kim N, et al. CT image conversion among different reconstruction kernels without a sinogram by using a convolutional neural network. *Korean J Radiol* 2019;20:295-303
24. Park S, Lee SM, Do KH, Lee JG, Bae W, Park H, et al. Deep learning algorithm for reducing CT slice thickness: effect on reproducibility of radiomic features in lung cancer. *Korean J Radiol* 2019;20:1431-1440
25. Ahn Y, Yoon JS, Lee SS, Suk HI, Son JH, Sung YS, et al. Deep learning algorithm for automated segmentation and volume measurement of the liver and spleen using portal venous phase computed tomography images. *Korean J Radiol* 2020;21:987-997
26. Hong JH, Park EA, Lee W, Ahn C, Kim JH. Incremental image noise reduction in coronary CT angiography using a deep learning-based technique with iterative reconstruction. *Korean J Radiol* 2020;21:1165-1177
27. Hwang EJ, Kim H, Yoon SH, Goo JM, Park CM. Implementation of a deep learning-based computer-aided detection system for the interpretation of chest radiographs in patients suspected for COVID-19. *Korean J Radiol* 2020;21:1150-1160
28. Koo HJ, Lee JG, Ko JY, Lee G, Kang JW, Kim YH, et al. Automated segmentation of left ventricular myocardium on cardiac computed tomography using deep learning. *Korean J Radiol* 2020;21:660-669
29. Park HJ, Shin Y, Park J, Kim H, Lee IS, Seo DW, et al. Development and validation of a deep learning system for segmentation of abdominal muscle and fat on computed tomography. *Korean J Radiol* 2020;21:88-100
30. Shin YJ, Chang W, Ye JC, Kang E, Oh DY, Lee YJ, et al. Low-dose abdominal CT using a deep learning-based denoising algorithm: a comparison with CT reconstructed with filtered back projection or iterative reconstruction algorithm. *Korean J Radiol* 2020;21:356-364
31. Weikert T, Noordtzij LA, Bremerich J, Stieltjes B, Parmar V, Cyriac 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;21:891-899
32. Zhou QQ, Wang J, Tang W, Hu ZC, Xia ZY, Li XS, et al. Automatic detection and classification of rib fractures on thoracic CT using convolutional neural network: accuracy and feasibility. *Korean J Radiol* 2020;21:869-879
33. Hwang HJ, Seo JB, Lee SM, Kim EY, Park B, Bae HJ, et al. Content-based image retrieval of chest CT with convolutional neural network for diffuse interstitial lung disease: performance assessment in three major idiopathic interstitial pneumonias. *Korean J Radiol* 2021;22:281-290
34. Kim JH, Yoon HJ, Lee E, Kim I, Cha YK, Bak SH. Validation of deep-learning image reconstruction for low-dose chest computed tomography scan: emphasis on image quality and noise. *Korean J Radiol* 2021;22:131-138
35. Kim K, Kim S, Han K, Bae H, Shin J, Lim JS. Diagnostic

performance of deep learning-based lesion detection algorithm in CT for detecting hepatic metastasis from colorectal cancer. *Korean J Radiol* 2021;22:912-921
36. Kim UH, Kim MY, Park EA, Lee W, Lim WH, Kim HL, et al. Deep learning-based algorithm for the detection and characterization of MRI safety of cardiac implantable electronic devices on chest radiographs. *Korean J Radiol* 2021;22:1918-1928
37. Lee JG, Kim H, Kang H, Koo HJ, Kang JW, Kim YH, et al. Fully automatic coronary calcium score software empowered by artificial intelligence technology: validation study using three CT cohorts. *Korean J Radiol* 2021;22:1764-1776
38. Lee KC, Lee KH, Kang CH, Ahn KS, Chung LY, Lee JJ, et al. Clinical validation of a deep learning-Based Hybrid (Greulich-Pyle and Modified Tanner-Whitehouse) method for bone age assessment. *Korean J Radiol* 2021;22:2017-2025
39. Park HS, Jeon K, Cho YJ, Kim SW, Lee SB, Choi G, et al. Diagnostic performance of a new convolutional neural network algorithm for detecting developmental dysplasia of the hip on anteroposterior radiographs. *Korean J Radiol* 2021;22:612-623
40. Purkayastha S, Xiao Y, Jiao Z, Thepumnoeysuk R, Halsey K, Wu J, et al. Machine learning-based prediction of COVID-19 severity and progression to critical illness using CT imaging and clinical data. *Korean J Radiol* 2021;22:1213-1224
41. Weikert T, Rapaka S, Grbic S, Re T, Chaganti S, Winkel DJ, et al. Prediction of patient management in COVID-19 using deep learning-based fully automated extraction of cardiothoracic CT metrics and laboratory findings. *Korean J Radiol* 2021;22:994-1004
42. Yan C, Lin J, Li H, Xu J, Zhang T, Chen H, et al. cycle-consistent generative adversarial network: effect on radiation dose reduction and image quality improvement in ultralow-dose CT for evaluation of pulmonary tuberculosis. *Korean J Radiol* 2021;22:983-993
43. Yang J, Chen Z, Liu W, Wang X, Ma S, Jin F, et al. Development of a malignancy potential binary prediction model based on deep learning for the mitotic count of local primary gastrointestinal stromal tumors. *Korean J Radiol* 2021;22:344-353
44. Yeoh H, Hong SH, Ahn C, Choi JY, Chae HD, Yoo HJ, et al. Deep learning algorithm for simultaneous noise reduction and edge sharpening in low-dose CT images: a pilot study using lumbar spine CT. *Korean J Radiol* 2021;22:1850-1857
45. Yoo SJ, Yoon SH, Lee JH, Kim KH, Choi HI, Park SJ, et al. Automated lung segmentation on chest computed tomography images with extensive lung parenchymal abnormalities using a deep neural network. *Korean J Radiol* 2021;22:476-488
46. Yu Y, Gao Y, Wei J, Liao F, Xiao Q, Zhang J, et al. A three-dimensional deep convolutional neural network for automatic segmentation and diameter measurement of type B aortic dissection. *Korean J Radiol* 2021;22:168-178
47. Bae K, Oh DY, Yun ID, Jeon KN. Bone suppression on chest radiographs for pulmonary nodule detection: comparison between a generative adversarial network and dual-energy subtraction. *Korean J Radiol* 2022;23:139-149
48. Chang S, Han K, Lee S, Yang YJ, Kim PK, Choi BW, et al. Automated measurement of native T1 and extracellular volume fraction in cardiac magnetic resonance imaging using a commercially available deep learning algorithm. *Korean J Radiol* 2022;23:1251-1259
49. Choi JW, Cho YJ, Ha JY, Lee YY, Koh SY, Seo JY, et al. deep learning-assisted diagnosis of pediatric skull fractures on plain radiographs. *Korean J Radiol* 2022;23:343-354
50. Kim YS, Jang MJ, Lee SH, Kim SY, Ha SM, Kwon BR, et al. Use of artificial intelligence for reducing unnecessary recalls at screening mammography: a simulation study. *Korean J Radiol* 2022;23:1241-1250
51. Lee JH, Kim KH, Lee EH, Ahn JS, Ryu JK, Park YM, et al. Improving the performance of radiologists using artificial intelligence-based detection support software for mammography: a multi-reader study. *Korean J Radiol* 2022;23:505-516
52. Otgonbaatar C, Ryu JK, Shin J, Woo JY, Seo JW, Shim H, et al. Improvement in image quality and visibility of coronary arteries, stents, and valve structures on CT angiography by deep learning reconstruction. *Korean J Radiol* 2022;23:1044-1054
53. Park HJ, Yoon JS, Lee SS, Suk HI, Park B, Sung YS, et al. Deep learning-based assessment of functional liver capacity using gadoxetic acid-enhanced hepatobiliary phase MRI. *Korean J Radiol* 2022;23:720-731
54. Park J, Shin J, Min IK, Bae H, Kim YE, Chung YE. Image quality and lesion detectability of lower-dose abdominopelvic CT obtained using deep learning image reconstruction. *Korean J Radiol* 2022;23:402-412
55. Park JH, Park I, Han K, Yoon J, Sim Y, Kim SJ, et al. Feasibility of deep learning-based analysis of auscultation for screening significant stenosis of native arteriovenous fistula for hemodialysis requiring angioplasty. *Korean J Radiol* 2022;23:949-958
56. Son W, Kim M, Hwang JY, Kim YW, Park C, Choo KS, et al. Comparison of a deep learning-based reconstruction algorithm with filtered back projection and iterative reconstruction algorithms for pediatric abdominopelvic CT. *Korean J Radiol* 2022;23:752-762
57. Yoo H, Kim EY, Kim H, Choi YR, Kim MY, Hwang SH, et al. Artificial intelligence-based identification of normal chest radiographs: a simulation study in a multicenter health screening cohort. *Korean J Radiol* 2022;23:1009-1018
58. Hwang EJ, Goo JM, Nam JG, Park CM, Hong KJ, Kim KH. Conventional versus artificial intelligence-assisted interpretation of chest radiographs in patients with acute respiratory symptoms in emergency department: a pragmatic randomized clinical trial. *Korean J Radiol* 2023;24:259-270
59. Lee SB, Hong Y, Cho YJ, Jeong D, Lee J, Yoon SH, et al. Deep learning-based computed tomography image standardization to improve generalizability of deep learning-based hepatic segmentation. *Korean J Radiol* 2023;24:294-304
60. Mnih V, Heess N, Graves A, Kavukcuoglu K. Recurrent models of visual attention. arXiv:1406.6247 [Preprint] 2014 [posted Jun 14 2014; cited Sep 15, 2023]. Available at: https://doi.

org/10.48550/arXiv.1406.6247
61. Clarivate. Journal Citation Reports(TM) for *Korean Journal of Radiology* [accessed on October 19, 2023]. Available at: https://jcr.clarivate.com/jcr-jp/journal-profile?journal=KOREAN%20J%20RADIOL&year=2022
62. Simago Journal & Country Rank. Simago Journal & Country Rank for *Korean Journal of Radiology* [accessed on October 19, 2023]. Available at: https://www.scimagojr.com/journalsearch.php?q=17255&tip=sid&exact=no
63. Soyer P. Agreement and observer variability. *Diagn Interv Imaging* 2018;99:53-54
64. Fernandez JC, Mounier L, Pachon C. *A model-based approach for robustness testing*. In: Khendek F, Dssouli R, eds. *Testing of Communicating Systems*. Berlin, Heidelberg: Springer, 2005:333-348
65. Joel MZ, Umrao S, Chang E, Choi R, Yang DX, Duncan JS, et al. Using adversarial images to assess the robustness of deep learning models trained on diagnostic images in oncology. *JCO Clin Cancer Inform* 2022;6:e2100170
66. Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. arXiv:1511.07122 [Preprint] 2016 [posted Nov 23 2015; revised Apr 30 2016; cited Sep 15, 2023]. Available at: https://doi.org/10.48550/arXiv.1511.07122
67. Camino R, Hammerschmidt CA, State R. *Working with deep generative models and tabular data imputation*. First Workshop on the Art of Learning with Missing Values (Artemiss) Hosted by the 37th International Conference on Machine Learning (ICML); 2020 Jul 12-18; Vienna, Austria
68. Radiological Society of North America. Scientific style guide: writing a manuscript for Radiology [accessed on October 16, 2023]. Available at: https://pubs.rsna.org/page/radiology/author-instructions/scientificediting
69. Silcox C, Dentzer S, Bates DW. AI-enabled clinical decision support software: a "Trust and Value Checklist" for clinicians. *NEJM Catal Innov Care Deliv* 2020;1