Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.243182

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Adherence to the Checklist for Artificial Intelligence in Medical Imaging (CLAIM): an umbrella review with a comprehensive two-level analysis

Burak Koçak[1]
Fadime Köse[1]
Ali Keleş[1]
Abdurrezzak Şendur[1]
İsmail Meşe[2]
Mehmet Karagülle[1]

[1]University of Health Sciences, Başakşehir Çam and Sakura City Hospital, Department of Radiology, İstanbul, Türkiye

[2]Üsküdar State Hospital, Department of Radiology, İstanbul, Türkiye

PURPOSE

To comprehensively assess Checklist for Artificial Intelligence in Medical Imaging (CLAIM) adherence in medical imaging artificial intelligence (AI) literature by aggregating data from previous systematic and non-systematic reviews.

METHODS

A systematic search of PubMed, Scopus, and Google Scholar identified reviews using the CLAIM to evaluate medical imaging AI studies. Reviews were analyzed at two levels: review level (33 reviews; 1,458 studies) and study level (421 unique studies from 15 reviews). The CLAIM adherence metrics (scores and compliance rates), baseline characteristics, factors influencing adherence, and critiques of the CLAIM were analyzed.

RESULTS

A review-level analysis of 26 reviews (874 studies) found a weighted mean CLAIM score of 25 [standard deviation (SD): 4] and a median of 26 [interquartile range (IQR): 8; 25th–75th percentiles: 20–28]. In a separate review-level analysis involving 18 reviews (993 studies), the weighted mean CLAIM compliance was 63% (SD: 11%), with a median of 66% (IQR: 4%; 25th–75th percentiles: 63%–67%). A study-level analysis of 421 unique studies published between 1997 and 2024 found a median CLAIM score of 26 (IQR: 6; 25th–75th percentiles: 23–29) and a median compliance of 68% (IQR: 16%; 25th–75th percentiles: 59%–75%). Adherence was independently associated with the journal impact factor quartile, publication year, and specific radiology subfields. After guideline publication, CLAIM compliance improved (P = 0.004). Multiple readers provided an evaluation in 85% (28/33) of reviews, but only 11% (3/28) included a reliability analysis. An item-wise evaluation identified 11 underreported items (missing in ≥50% of studies). Among the 10 identified critiques, the most common were item inapplicability to diverse study types and subjective interpretations of fulfillment.

CONCLUSION

Our two-level analysis revealed considerable reporting gaps, underreported items, factors related to adherence, and common CLAIM critiques, providing actionable insights for researchers and journals to improve transparency, reproducibility, and reporting quality in AI studies.

CLINICAL SIGNIFICANCE

By combining data from systematic and non-systematic reviews on CLAIM adherence, our comprehensive findings may serve as targets to help researchers and journals improve transparency, reproducibility, and reporting quality in AI studies.

KEYWORDS

Artificial intelligence, machine learning, checklist, diagnostic imaging, radiology

Corresponding author: Burak Koçak

E-mail: drburakkocak@gmail.com

Received 11 December 2024; revision requested 04 January 2025; last revision received 19 January 2025; accepted 22 January 2025.

Epub: 10.02.2025

Publication date:

DOI: 10.4274/dir.2025.243182

You may cite this article as: Koçak B, Köse F, Keleş A, Şendur A, Meşe İ, Karagülle M. Adherence to the Checklist for Artificial Intelligence in Medical Imaging (CLAIM): an umbrella review with a comprehensive two-level analysis. *Diagn Interv Radiol.* 10 February 2025 DOI: 10.4274/dir.2025.243182 [Epub Ahead of Print].

With the exponential increase in artificial intelligence (AI) publications related to medical imaging,[1] ensuring transparency and reproducibility has become crucial for advancing the field and integrating AI into clinical practice.[2-4] To address these needs, various AI-focused reporting guidelines have been introduced,[5-7] one of which is the Checklist for Artificial Intelligence in Medical Imaging (CLAIM).[8] Published in March 2020, the CLAIM was designed to improve reporting clarity and scientific communication in medical imaging AI.[8] Inspired by the Standards for Reporting of Diagnostic Accuracy Studies (STARD) guidelines,[9] the original 2020 version of the CLAIM featured a 42-item checklist to help authors and reviewers achieve clear, comprehensive, and reproducible reporting in AI studies. In May 2024, an updated CLAIM was published following a formal Delphi process, refining the checklist to 44 items to address new challenges and developments while retaining the original structure.[10] The update included refinements to terminology and revisions to some items. The CLAIM is part of the EQUATOR network, a central hub for reporting guidelines.[11]

Since its release, the CLAIM has gained widespread attention across multiple medical specialties involving imaging and AI, with over 850 citations in Google Scholar as of January 2025. Despite its popularity, assessments of CLAIM adherence remain highly variable,[12-14] often with particular focus on specific diseases,[15-18] techniques,[19-21] or individual journals.[22] A comprehensive assessment of CLAIM adherence across these diverse studies is notably lacking. Such an analysis, previously applied to frameworks such as the Radiomics Quality Score (RQS),[23] would reveal the CLAIM's overall adherence patterns, highlight underreported items, and provide guidance for future revisions beyond the 2024 CLAIM update,[10] along with the development of new, alternative AI checklists.

Main points

- To our knowledge, no prior research has synthesized data from published reviews on Checklist for Artificial Intelligence in Medical Imaging (CLAIM) adherence, leaving a gap in providing a comprehensive overview independent of disease, technique, or journal.
- Our two-level analysis identified significant reporting gaps in the medical imaging artificial intelligence literature, with a third of CLAIM items omitted, on average.
- Eleven specific CLAIM items were identified as being consistently underreported in the majority of studies, highlighting critical areas for improvement.
- Factors such as the publication year, journal impact quartile, and the radiology subfield influenced CLAIM adherence.
- Reviews assessing CLAIM adherence exhibited variability in their methodologies, with some using scoring and others focusing on compliance, leading to inconsistencies in evaluation and reporting.

This study aims to comprehensively assess CLAIM adherence in the medical imaging AI literature published to date using a two-level approach: review level and study level. The review-level analysis aggregates data from previous systematic and non-systematic reviews, whereas the study-level analysis examines unique individual papers within these reviews, mostly focusing on checklist items. Furthermore, factors influencing high or low CLAIM adherence are examined at the study level. Finally, critiques of the CLAIM guidelines are systematically analyzed across eligible reviews for both levels.

Methods

Literature search and screening

A literature search was conducted through PubMed, Scopus, and Google Scholar to identify reviews on the application of the CLAIM[8] using the syntax "Checklist for Artificial Intelligence in Medical Imaging." The final search was performed on August 6, 2024. Since the search syntax was simple, we did not use advanced database features to target specific fields (e.g., title, abstract, or keywords). Instead, we used the general search box, which typically searches across all fields in the database entries.

For Google Scholar, the first 100 results were screened based on the filter setting "relevance," whereas all entries were reviewed in the other two databases. Google Scholar can provide valuable additions to systematic reviews, even when screening is limited to the top 100 results.[24] Because its "relevance"-based ranking typically prioritizes the most pertinent articles, this approach was chosen to manage the large volume of results often retrieved from Google Scholar, many of which include duplicates or less relevant entries. Notably, Google Scholar was treated as a supplementary source to mitigate the risk of missing key papers, complementing the more comprehensive searches conducted in PubMed and Scopus, where all entries were reviewed.

Three readers (F.K., A.K., and A.S.; all 3rd- or 4th-year radiology residents) initially screened all records to identify review articles evaluating medical imaging AI studies using the CLAIM (2020 version).[8] Records were excluded if they lacked a CLAIM evaluation (2020 version),[8] full-text access, and relevance to medical imaging; relied on self-reported data; or had significant overlap with another study. Each reader cross-checked another reader's results.

Duplicates were removed using Zotero software. The full-text articles and available supplements were downloaded for evaluation by the same three readers, who divided the workload equally. For articles where full-text access was unavailable through our institutional libraries, we tried to reach out directly to the authors to request access.

Eligibility

After the initial screening, articles were evaluated for eligibility by the same three readers under the supervision of a radiology specialist experienced in informatics and AI (B.K.). For the review-level analysis, reviews with adequate adherence data on the 42-item CLAIM were included; those with incomplete or unclear data were excluded. For the study-level analysis, only reviews with 42-item CLAIM data for each study (i.e., a completed checklist for each study) were included. Duplicate and retracted studies, along with the studies with unclear references to their source articles, were removed. Papers using a modified 42-item CLAIM with subsections that retained the main framework were included in the study-level analysis but excluded from the review-level analysis unless CLAIM adherence could be evaluated at that level.

Analyzing data at the individual study level was crucial to gain item-level insights as well as several other baseline characteristics, as this level of granularity could not have been achieved through a review-level-only analysis. Although we acknowledge the potential limitations of using a highly selected sample, this approach was necessary to address the study's objectives and provide meaningful insights at the desired level of detail.

Data extraction

For the review-level analysis, data extraction was initially performed by a radiology specialist experienced in informatics and AI (B.K.) and was subsequently confirmed by another radiology specialist (M.K.). Extracted data included the review's scope, radiology

subfield, number of studies (or evaluations) in the reviews, online publication year, number of readers, reader independence, decision-making methods, reproducibility analysis, consideration of non-applicable (n/a) items in the adherence evaluation, CLAIM adherence evaluation method, and source of the CLAIM evaluation.

For the study-level analysis, the three radiology residents independently extracted and cross-checked the data. The cross-checking was performed by having the readers review and validate one another's work. In cases of disagreement, an experienced reader (B.K.) was consulted to resolve the issue. Extracted information included the journal name, publication year, publication type, journal scope and focus, radiology subfield (expanded from the review-level data), journal's h5-index (from Google Scholar Metrics), 2023 impact factor quartile (2024 release; Journal Citation Reports, Clarivate Analytics, Web of Science Group), and CLAIM adherence by item.

Full-text articles, including the text, figures, tables, and supplements, were reviewed to identify adherence data, including item-specific CLAIM data, organized according to the original item order, if necessary. For adherence data sourced from the reviews, only studies with a clear source attribution were included. In cases of multiple rater evaluations, consensus data were prioritized; if unavailable, one evaluation (the first) was selected. In the study-level analysis, only one assessment per study was included when multiple pipelines were assessed, whereas all assessments were considered in the review-level analysis, which are referred to as "studies" in this research. For studies using a modified CLAIM with subsections within a 42-item framework, an item was considered reported if ≥50% of its subitems were positively evaluated. Partially reported items were classified as reported, in alignment with the common standard checklist format (i.e., reported, not reported, and not applicable).

Two radiology specialists with experience in informatics and AI (B.K. and İ.M.) evaluated the review papers in both the review-level and study-level analyses for critiques about the CLAIM. The PDFs were then screened using Google's NotebookLM tool, with various targeted prompts to identify additional critiques and to minimize the risk of missing important ones. The results from this additional screening were double-checked by both readers, verified against their sources, and integrated with the initial human evaluation findings.

Adherence metrics

This study applied two commonly used CLAIM adherence metrics: the CLAIM score and CLAIM compliance. The CLAIM score represents the total number of reported items, whereas CLAIM compliance is calculated as the percentage of reported items relative to the total applicable CLAIM items.

For the study-level analysis, these two metrics were calculated directly from the extracted item-level data. In the review-level analysis, metrics were extracted as a mean and used as reported when directly provided; if not, they were derived from tables, figures, or supplementary files where possible, converted from the median and interquartile range (IQR), if necessary, according to the methods proposed by Luo et al.[25] and Wan et al.[26], or computed as weighted combinations when presented by category.

Statistical analysis

Statistical analysis was conducted using R (main packages: ggstatsplot and Hmisc) and JASP (version 0.19.1; Apple Silicon). Descriptive statistics, including frequency, percentage, mean, standard deviation (SD), median, IQR, and 25th–75th percentiles, were reported based on variable distribution. In the review-level analysis, adherence metrics were weighted by the number of studies or evaluations using the "Hmisc" R package and presented using both the mean and median without considering statistical normality. For the study-level data, normality was tested with the Shapiro–Wilk test, and the associated statistical results are presented accordingly. In addition, differences between continuous variables were assessed using the Mann–Whitney U test or Student's t-test based on distribution. The Kruskal–Wallis test was applied to compare multiple categories, with Dunn's post-hoc tests and the Bonferroni correction. Correlations were assessed with Spearman's rho. Univariable and multivariable logistic regression was performed to identify the potential factors related to high and low CLAIM adherence metrics according to the median. No multiplicity correction was performed in the logistic regression analyses due to the exploratory nature of the study. Statistical significance was set at $P < 0.05$.

Results

Literature search

Figure 1 summarizes the eligibility process. Finally, 33 eligible reviews encompassing 1,458 study evaluations were included in the review-level analysis. For the study-level analysis, 15 reviews (13 from the previous set and 2 additional reviews) were included, covering 421 unique eligible studies. In total, 35 reviews met the eligibility criteria for both levels of analysis (Table 1).[12-22,27-50] The final dataset used in this study is publicly available from the Open Science Framework and can be accessed via the following link: https://osf.io/rx67y/

Baseline characteristics of papers eligible for the review-level analysis

The baseline characteristics of the 33 papers included in the review-level analysis are summarized in Table 2.

Multiple readers conducted CLAIM evaluations in 85% of reviews (28/33), with most assessments (79%, 22/28) performed independently and finalized by consensus (82%, 23/28). A reliability analysis was included in only a few multi-reader studies (11%, 3/28). One study reported an intraclass correlation coefficient (ICC) above 0.87 for inter-observer reliability across task categories.[46] Another study found an ICC of 0.815 for inter-observer reliability, with varying kappa values for individual items.[14] A third study reported an intra-observer repeatability coefficient of 0.22, which was lower and better than that of other checklists evaluated, except one.[31]

Figure 2 highlights the consideration of item applicability in the included reviews, along with the resultant metrics from this study. Regarding CLAIM adherence, 55% (18/33) of reviews considered the applicability of items, allowing for the calculation of a CLAIM compliance metric. For approximately 79% (26/33) of the reviews, appropriate data to calculate CLAIM scores were available, although the origin of the scores varied, with only 36% (12/33) providing direct reports.

Adherence based on the review-level analysis

Among the 26 reviews with available CLAIM scores, encompassing 874 studies, the weighted mean CLAIM score was 25 (SD: 4), and the weighted median was 26 (IQR: 8; 25th–75th percentiles: 20–28). For the 18 reviews providing CLAIM compliance data, covering 993 studies, the weighted mean CLAIM compliance was 63% (SD: 11%), with

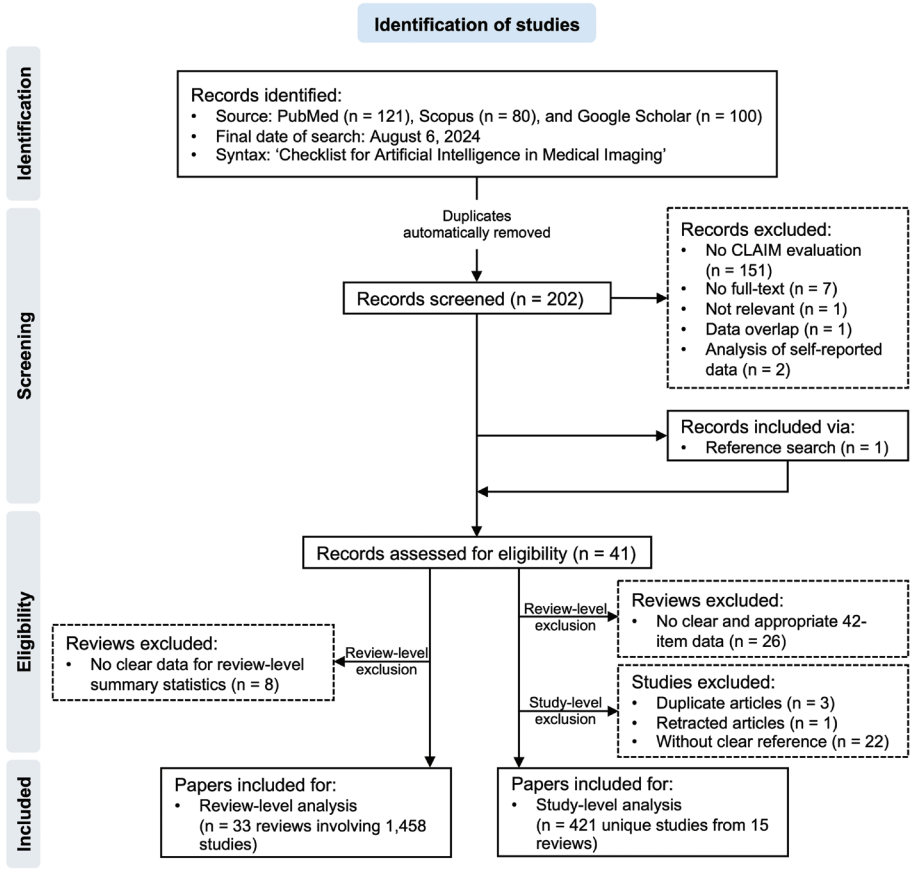

Figure 1. Identification of eligible studies for the review- and study-level analyses. CLAIM, Checklist for Artificial Intelligence in Medical Imaging.

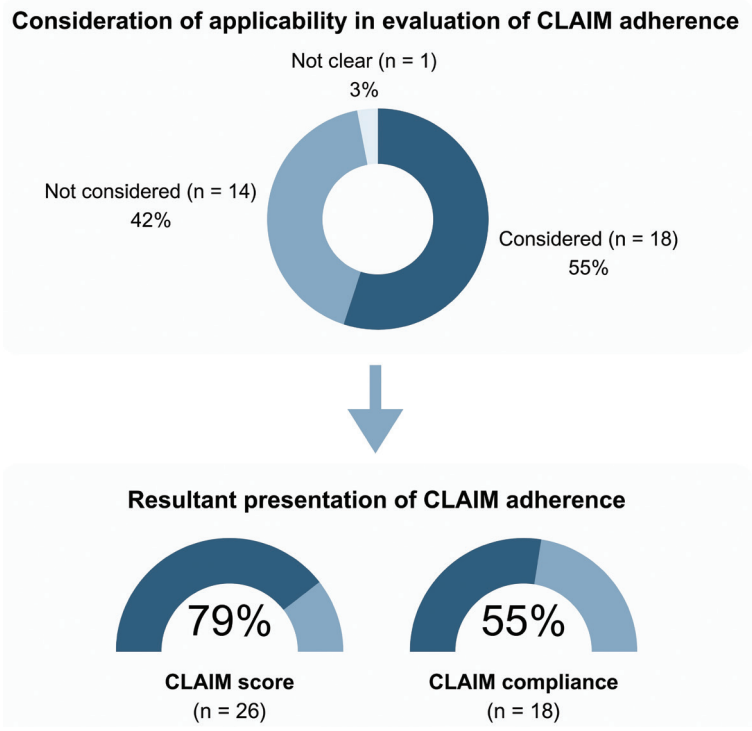

Figure 2. Consideration of item applicability and resultant CLAIM adherence metrics in the review-level analysis, emphasizing the methodological variability among reviews evaluating CLAIM adherence. CLAIM, Checklist for Artificial Intelligence in Medical Imaging.

a weighted median of 66% (IQR: 4%; 25th–75th percentiles: 63%–67%).

Baseline characteristics of papers eligible for the study-level analysis

The baseline characteristics of the papers included in the study-level analysis are summarized in Table 3. Publication dates ranged from 1997 to 2024.

Adherence based on the study-level analysis

In the study-level analysis of 421 unique studies, the median CLAIM score was 26 (IQR: 6; 25th–75th percentiles: 23–29), and the median CLAIM compliance was 68% (IQR: 16%; 25th–75th percentiles: 59%–75%). Notably, 11% of the studies (47/421) had a CLAIM score of <21 (i.e., 50% of 42), whereas 10% (40/421) reported a CLAIM compliance of <50%.

Figure 3 illustrates the median CLAIM scores and compliance by journal and publication volume. Among the top 10 journals by publication volume, *Radiology* had the highest median CLAIM score and compliance rate.

Table 4 presents the results from the univariable and multivariable logistic regression analyses to identify factors linked to high and low CLAIM adherence. In the univariable analysis, the publication year, specific radiology subfields, journal h5-index, and certain impact factor quartiles were associated with the CLAIM score or compliance. In the multivariable analysis, the publication year and impact factor quartile emerged as independent predictors of the CLAIM score and compliance. Specifically, publishing in a first quartile (Q1) journal independently predicted higher CLAIM scores and compliance, whereas second quartile (Q2) journals were associated with higher CLAIM compliance. Certain radiology subfields were additional independent predictors of the CLAIM score.

Figure 4a, b illustrate the correlation between the publication year and CLAIM adherence. Although the CLAIM score did not significantly correlate with the publication year (rho: 0.076, $P = 0.117$), CLAIM compliance showed a weak but significant positive correlation (rho: 0.119, $P = 0.015$). Although the CLAIM score did not significantly differ between the pre- and post-CLAIM guideline publication periods ($P = 0.153$), CLAIM compliance was higher post-publication ($P = 0.004$) (Figure 4c, d). However, neither the CLAIM score (rho: −0.027, $P = 0.697$)

Table 1. Reviews included in the analyses, detailing the authors, year, journal abbreviation, radiology subfield, and the number of papers or evaluations included in the review- and study-level analyses

Authors (online publication year)	Journals	Radiology subfield	No. of papers or evaluations[1]	
			Review level	Study level
Abdulaal et al.[15] (2024)	Front Radiol	Chest	5	5
Alabed et al.[19] (2022)	Front Cardiovasc Med	Cardiovascular	209	n/a
Alipour et al.[16] (2023)	Diagnostics (Basel)	Musculoskeletal	8	n/a
Assadi et al.[27] (2022)	Medicina (Kaunas)	Cardiovascular	5	5
Bedrikovetski et al.[28] (2022)	Eur J Radiol	General or multi-system	24	24
Belue et al.[12] (2022)	J Am Coll Radiol	Genitourinary	53	n/a
Belue and Turkbey[29] (2022)	Eur Radiol Exp	Genitourinary	47	n/a
Bhandari et al.[13] (2023)	Neuroradiology	Neuro	138	n/a
Bleker et al.[30] (2022)	Life (Basel)	Genitourinary	4	4
Cerdá-Alberich et al.[31] (2023)	Insights Imaging	General or multi-system	10	9
Dagher et al.[32] (2024)	J Neuroimaging	Neuro	6	n/a
Hardacre et al.[33] (2021)	Br J Radiol	Cardiovascular	3	3
Hickman et al.[34] (2021)	Radiology	Breast	14	n/a
Hu et al.[35] (2022)	Neuroradiology	Neuro	19	n/a
Hwang et al.[36] (2024)	Radiol Artif Intell	Chest	14	n/a
Jia et al.[37] (2022)	Eur J Radiol Open	Chest	19	7
Karabacak et al.[20] (2022)	Acta Radiol	Neuro	5	n/a
Karabacak et al.[38] (2022)	Quant Imaging Med Surg	Neuro	4	n/a
Kim et al.[22] (2023)	Korean J Radiol	General or multi-system	38	n/a
Kouli et al.[21] (2022)	Neurooncol Adv	Neuro	234	222
Lans et al.[39] (2022)	Artif Intell Med	Musculoskeletal	91	n/a
Le et al.[40] (2021)	Appl Sci	Dental	6	6
O'Shea et al.[41] (2021)	Eur Radiol	General or multi-system	186	n/a
Ozkara et al.[18] (2023)	Cancers (Basel)	Neuro	25	n/a
Raj et al.[42] (2024)	Indian J Orthop	Musculoskeletal	5	n/a
Roberts et al.[43] (2021)	Nat Mach Intell	Chest	37	37
Roest et al.[44] (2022)	Life (Basel)	Genitourinary	8	n/a
Si et al.[14] (2021)	Eur Radiol	Musculoskeletal	36	36
Sivanesan et al.[45] (2022)	Can Assoc Radiol J	General or multi-system	100	n/a
Sushentsev et al.[17] (2022)	Insights Imaging	Genitourinary	5	5
Tsang et al.[46] (2023)	Jpn J Radiol	Pediatric	21	21
Wang et al.[48] (2023)	Radiother Oncol	Neuro	42	n/a
Wang et al.[47] (2024)	Radiother Oncol	Chest	37	n/a
Zhong et al.[49] (2022)	Insights Imaging	Musculoskeletal	n/a	28
Zhong et al.[50] (2023)	J Orthop Surg Res	Musculoskeletal	n/a	9

[1]Values represent the total number of studies or evaluations (i.e., pipelines) included in our analysis after applying the eligibility criteria and therefore may not correspond exactly to the total number of studies reported in the respective papers. n/a, not available.

nor compliance (rho: –0.062, $P = 0.365$) was statistically significantly correlated with the publication year after the CLAIM guideline publication in 2020.

The CLAIM scores and compliance varied significantly across radiology subfields ($P < 0.001$ for both), with post-hoc pairwise comparisons showing that the cardiovascular subfield had consistently distinct results compared with others (Figure 5).

The CLAIM scores and compliance also differed by impact factor quartile ($P < 0.001$ for CLAIM score; $P = 0.002$ for CLAIM compliance) (Figure 6). The post-hoc analysis revealed that journals in Q1 and Q2 had significantly higher CLAIM scores than non-Web of Science indexed journals or publication platforms. However, CLAIM compliance did not show significant pairwise differences across quartiles.

Moreover, the CLAIM scores and compliance were not statistically significantly different among different publication types, such as journal articles, pre-prints, and conference papers ($P > 0.05$).

The item-wise CLAIM adherence is presented in Figure 7. Notably, three items were mostly n/a in ≥50% of the papers: item#10 (selection of data subsets, if applicable), item#21 (the level at which partitions are disjoint, e.g., image, study, patient, institution), and item#27 (ensemble techniques, if applicable).

Considering the applicability of the items, the following 11 items were not reported in ≥50% of the papers (i.e., compliance of <50%): item#12 (de-identification methods), item#13 (how missing data were handled), item#19 (intended sample size and how it was determined), item#29 (statistical measures of significance and uncertainty, e.g., confidence intervals), item#31 (methods for explainability or interpretability and how they were validated), item#33 (flow of participants or cases, using a diagram to indicate

No.	Publication Platform	Frequency	CLAIM Score	CLAIM Compliance
1	International journal of imaging systems and technology	21	24	62
2	European radiology	11	28	71
3	IEEE Access	11	27	68
4	Insights into imaging	10	27	63
5	PloS one	10	26	65
6	Arxiv	9	25	74
7	Radiology	9	32	78
8	Computerized medical imaging and graphics	9	27	68
9	Computer methods and programs in biomedicine	9	27	69
10	Computers in biology and medicine	8	29	73
11	IEEE transactions on medical imaging	7	27	72
12	Journal of magnetic resonance imaging	6	29	74
13	Medical image analysis	6	27	73
14	Scientific reports	6	30	74
15	Medical physics	5	26	67
16	European journal of radiology	5	26	67
17	Journal of digital imaging	5	30	76
18	International journal of computer assisted radiology and surgery	5	28	72
19	American journal of neuroradiology	5	29	71
20	Neurocomputing	5	29	74
21	Biocybernetics and biomedical engineering	5	25	64
22	Multimedia tools and applications	5	26	67
23	Biomedical signal processing and control	5	26	67
24	Frontiers in oncology	5	35	83
25	MedRxiv	5	24	74

a

No.	Publication Platform	Frequency	CLAIM Score	CLAIM Compliance
1	Circulation	1	41	98
2	Journal of the American Heart Association	1	40	95
3	European heart journal cardiovascular imaging	1	37	95
4	Heart	1	39	93
5	JACC cardiovascular imaging	1	39	93
6	Journal of medical Internet research	1	29	91
7	European respiratory journal	1	30	88
8	Neuro-oncology	1	34	87
9	EBioMedicine	1	36	86
10	European radiology experimental	3	35	85
11	Physica medica	1	28	85
12	Neuroradiology	1	33	85
13	Korean journal of radiology	1	35	85
14	Experimental and therapeutic medicine	1	27	84
15	Cancer imaging	2	34	84
16	Frontiers in oncology	5	35	83
17	Radiology: artificial intelligence	3	35	83
18	BioMed research international	2	35	83
19	International journal of medical informatics	1	25	83
20	Chinese journal of radiology	1	35	83
21	Quantitative imaging in medicine and surgery	1	35	83
22	Clinical neuroradiology	1	32	82
23	The lancet digital health	1	32	82
24	NeuroImage	1	32	80
25	PLoS medicine	1	32	80

b

Figure 3. Tabulated bar charts for the study-level analysis of the median CLAIM score and compliance by journal, sorted by publication frequency **(a)** and CLAIM compliance **(b).** CLAIM, Checklist for Artificial Intelligence in Medical Imaging.

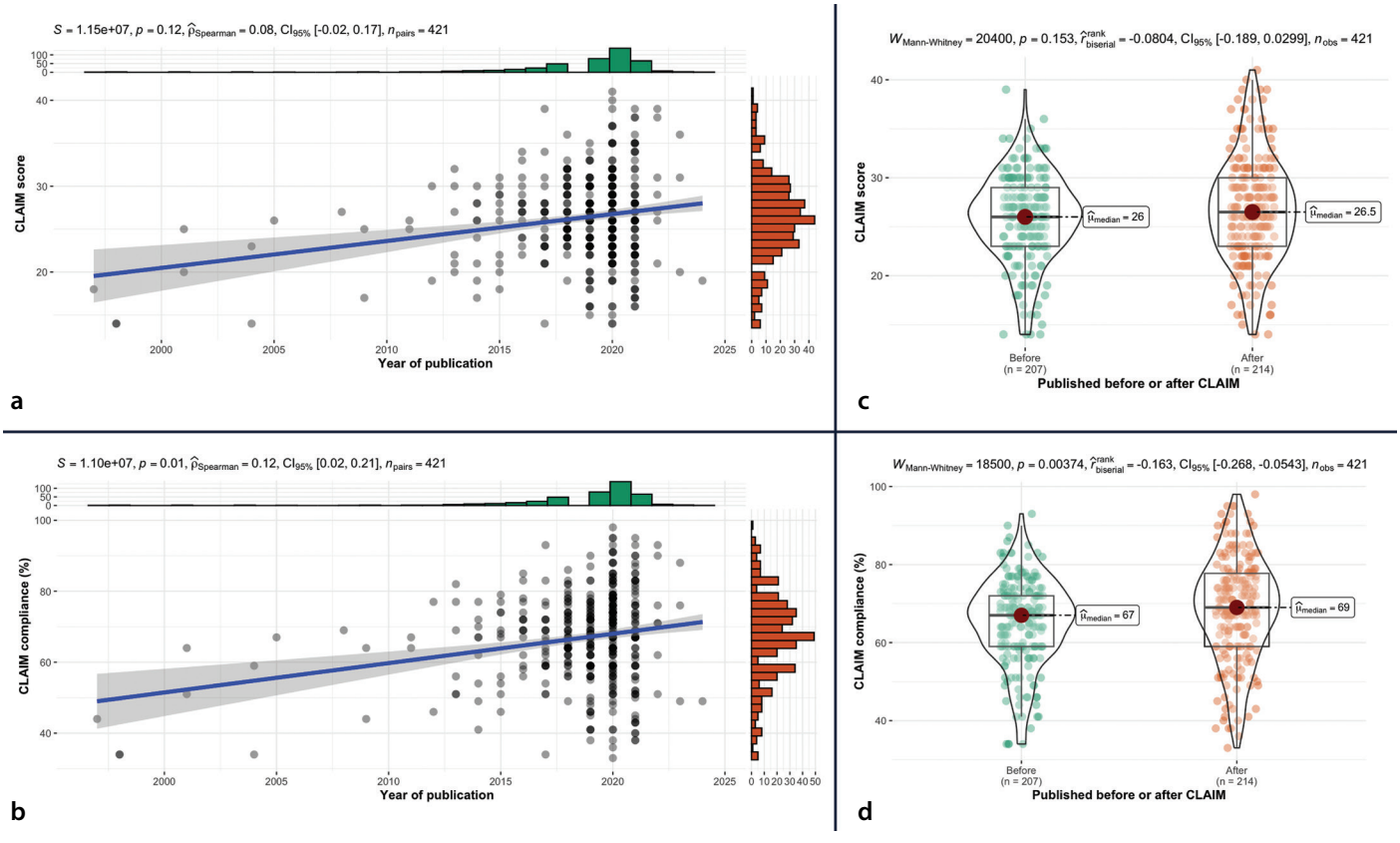

Figure 4. Study-level analysis of the publication year, CLAIM score, and compliance. Scatterplots with marginal distributions showing the correlation between the publication year and CLAIM score **(a)** and compliance **(b)**. Combined box and violin plots illustrating the CLAIM score **(c)** and compliance **(d)** in relation to the release of the CLAIM guidelines in 2020. CLAIM, Checklist for Artificial Intelligence in Medical Imaging; CI, confidence interval.

inclusion and exclusion), item#34 (demographic and clinical characteristics of cases in each partition), item#36 (estimates of diagnostic accuracy and their precision), item#37 (failure analysis of incorrectly classified cases), item#40 (registration number and name of registry), and item#41 (where the full study protocol can be accessed). Figure 8 further highlights the above-mentioned 11 items categorized into three domains: data handling and description, model evaluation and performance, and open science.

The item-wise correlation results for reporting status and year are presented in Table 5, according to pre- and post-publication and post publication of the CLAIM. Considering the entire period, a positive weak-to-moderate and statistically significant reporting trend (rho ≥0.2) was observed for item#19 (intended sample size and how it was determined), item#21 (level at which partitions are disjoint), item#31 (methods for explainability or interpretability and how they were validated), item#33 (flow of participants or cases, using a diagram to indicate inclusion and exclusion), and item#42 (sources of funding and other support; role of funders). Moreover, a negative weak-to-moderate reporting trend (rho ≤−0.2) was observed for item#11 (definitions of data elements, with references to common data elements), item#15 (rationale for choosing the reference standard), item#17 (annotation tools), item#18 (measurement of inter- and intra-rater variability), and item#39 (implications for practice, including the intended use and/or clinical role). Considering the post-publication period, a positive weak-to-moderate reporting trend (rho ≥0.2) was observed in item#10 (selection of data subsets), item#19 (intended sample size and how it was determined), and item#33 (flow of participants or cases, using a diagram to indicate inclusion and exclusion). In addition, a negative weak-to-moderate reporting trend (rho ≤−0.2) was observed for item#9 (data pre-processing steps) and item#39 (implications for practice, including the intended use and/or clinical role).

Table 2. Baseline characteristics of eligible papers included in the review-level analysis

Characteristic	Sub-category	Value
Scope, count (%)	Broad (AI, ML, or deep learning)	22 (67%)
	Deep learning	9 (27%)
	Radiomics	2 (6%)
Radiology subfield, count (%)	Neuro	8 (24%)
	Chest	5 (15%)
	Genitourinary	5 (15%)
	General or multi-system	5 (15%)
	Musculoskeletal	4 (12%)
	Cardiovascular	3 (9%)
	Pediatric	1 (3%)
	Breast	1 (3%)
	Dental	1 (3%)
Number of papers within reviews, median (IQR; 25th–75th percentiles)	-	19 (36; 6–42)
Publication year (online), count (%)	2021	6 (18%)
	2022	15 (45%)
	2023	7 (21%)
	2024	5 (15%)
Number of readers, count (%)	Multiple	28 (85%)
	Single	4 (12%)
	Not clear	1 (3%)
Dependence of reading, count (%)	Independent	22 (67%)
	Not clear	6 (18%)
	Not applicable	5 (15%)
Final decision of reading, count (%)	Consensus	23 (70%)
	Not clear	5 (15%)
	Not applicable	5 (15%)
Reliability analysis, count (%)	No	25 (76%)
	Not applicable	5 (15%)
	Yes	3 (9%)
Source of CLAIM evaluation, count (%)	As reported	12 (36%)
	Calculated from table or figure data	15 (45%)
	As reported + calculated from table or figure	2 (6%)
	As reported with median–mean conversion	3 (9%)
	As reported with a weighted combination of different categories	1 (3%)

Percentages may not total 100% due to rounding. IQR, interquartile range; CLAIM, Checklist for Artificial Intelligence in Medical Imaging; AI, artificial intelligence; ML, machine learning.

Critiques in reviews eligible for the entire study

In analyzing the 35 reviews that applied the CLAIM, we identified 10 key critiques, which we organized into 7 categories: fulfillment, applicability, feasibility and practicality, structure, interpretation, relative importance, and scoring. The most common critique was the inapplicability of certain items to all study types. Another frequent issue was the subjective nature of deciding whether an item was sufficiently reported. Table 6 presents all the critiques along with their representative source articles.

Discussion

Main findings and related implications

This study comprehensively evaluated CLAIM adherence in the medical imaging AI literature through a two-level approach: review- and study-level analyses. Considering both analyses, on average, one-third of CLAIM items were inadequately reported, indicating room for improvement in adhering to reporting guidelines. Since adherence was independently assessed rather than self-reported, efforts to improve compliance should focus on improving awareness and engagement among researchers in terms of transparent reporting practices through guidelines. Notwithstanding their

well-known benefits,[51] recent meta-research shows that radiology, nuclear medicine, and medical imaging journals rarely mandate AI-specific guidelines, despite the CLAIM being the most recommended.[52,53] Journals can actively endorse and promote the CLAIM[8] and its updates[10] to improve reporting quality and transparency while ensuring proper checklist usage with auditing practices.[54,55]

Our correlation analysis revealed a very weak but positive trend between CLAIM compliance and publication year. Although compliance was higher in the post-publication period, the trend was not statistically significant. Long-term follow-up studies on checklists such as STARD have demonstrated slow but significant improvements in research reporting quality over time.[56] Although a similar trend was observed in our analysis, more time and data are needed to better understand this progression and assess the CLAIM's true impact.

We observed that adherence assessments in reviews often lacked consistency due to the absence of standardized methods. We identified two primary approaches, the CLAIM score and CLAIM compliance (%), differing by item applicability. To improve comparability and fairness in the evaluation of adherence, we strongly recommend prioritizing the CLAIM compliance rate over the CLAIM score in future evaluations. The compliance rate accounts for the applicability of individual items, which can vary between studies, thereby providing a more accurate and equitable assessment. Moreover, this approach could be formally recommended or mandated by the developers in future versions of the CLAIM to ensure consistent and standardized adherence evaluations.

Publication year, impact factor quartile, and radiology subfields were key independent predictors of high or low CLAIM adherence. Studies in higher-impact journals (Q1 and Q2) showed stronger adherence, underscoring their role in setting transparent reporting standards and enabling rigorous peer review. However, it should be acknowledged that high-quality research can also be published in lower-impact journals, and high-impact journals are not immune to poor-quality research. Factors contributing to stronger adherence in higher-impact journals may include stricter editorial and peer-review processes, greater visibility of reporting guidelines in these journals, and, potentially, a higher familiarity of authors with these standards. In this respect, encouraging CLAIM adoption, particularly in lower-impact journals, could help enhance reporting transparency and reproducibility. It is important to note, however, that these observations are based on assumptions and warrant further investigation.

In addition, certain subfields, such as cardiovascular imaging, exhibited unique adherence patterns, reflecting differences in the maturity of AI reporting practices. These findings may indicate the need for specific strategies to improve CLAIM adherence across diverse medical imaging subfields and ensure consistent reporting standards throughout the discipline. Further research may be required to investigate whether unique adherence patterns in certain subfields, such as cardiovascular imaging, could

Table 3. Baseline characteristics of eligible papers included in the study-level analysis

Variable	Category	Value
Radiology subfield, count (%)	Neuro	222 (53%)
	Musculoskeletal	73 (17%)
	Chest	49 (12%)
	General or multi-system	33 (8%)
	Pediatric	21 (5%)
	Genitourinary	9 (2%)
	Cardiovascular	8 (2%)
	Dental	6 (1%)
Publication type, count (%)	Journal article	403 (96%)
	Preprint	14 (3%)
	Conference paper	4 (1%)
Scope of journals, count (%)	Radiology or imaging-related	170 (40%)
	No	251 (60%)
Focus of journals, count (%)	AI-focused	14 (3%)
	No	407 (97%)
h-5 index of journal, median (IQR; 25th–75th percentiles)	-	67 (70; 44–113)
Impact factor quartile, count (%)	Q1	222 (53%)
	Q2	116 (28%)
	Q3	29 (7%)
	Q4	17 (4%)
	No	37 (9%)
Top 10 most frequent publication platform, count (%)	International journal of imaging systems and technology	21 (5%)
	European radiology	11 (3%)
	IEEE access	11 (3%)
	Insights into imaging	10 (2%)
	PloS one	10 (2%)
	Arxiv	9 (2%)
	Radiology	9 (2%)
	Computerized medical imaging and graphics	9 (2%)
	Computer methods and programs in biomedicine	9 (2%)
	Computers in biology and medicine	8 (2%)

IQR, interquartile range; AI, artificial intelligence.

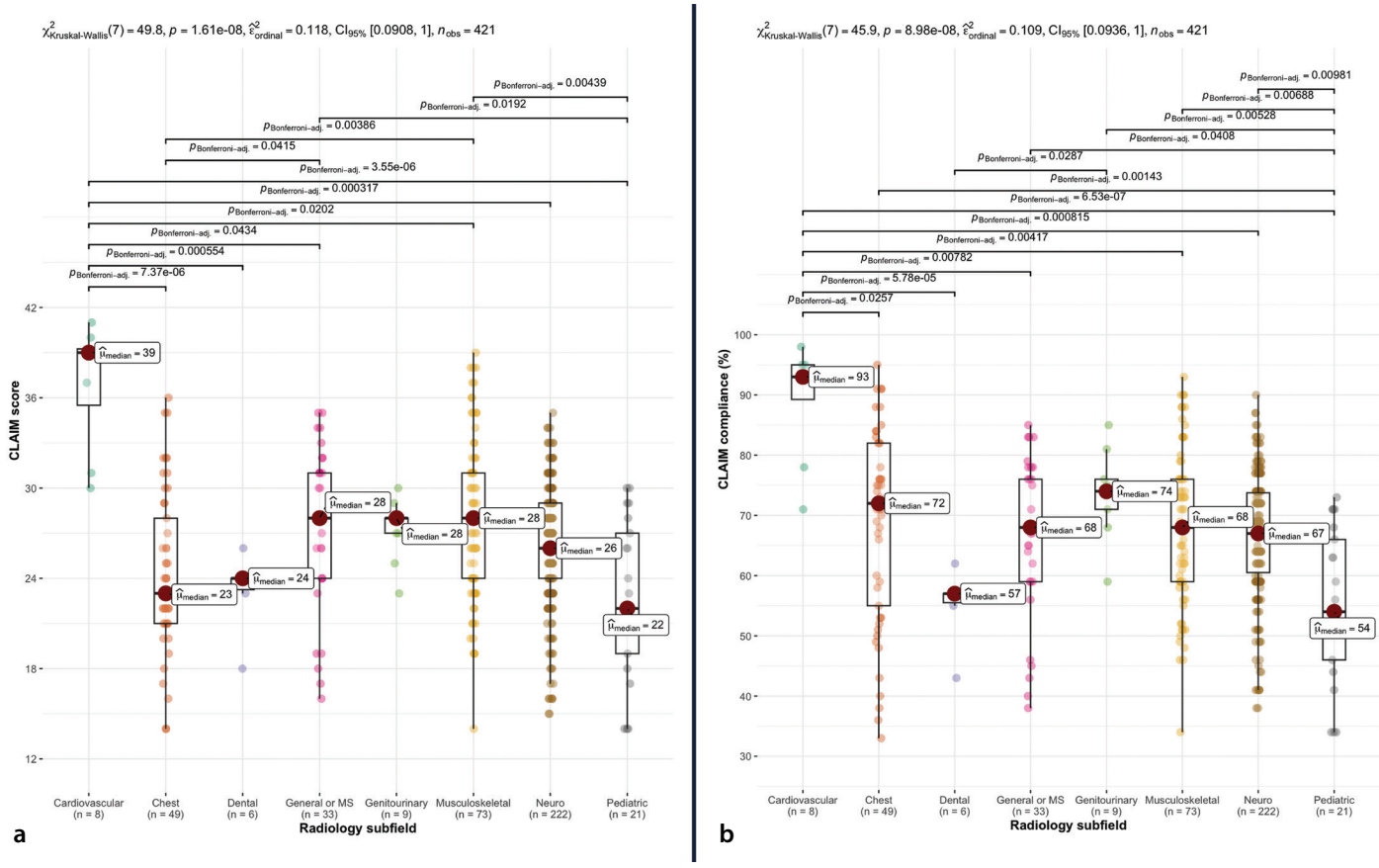

Figure 5. Box plots for the study-level analysis of the CLAIM score **(a)** and compliance **(b)** by radiology subfield, with pairwise comparisons. The Kruskal–Wallis test showed statistically significant differences across all categories in both analyses **(a, b)**. Only statistically significant pairwise comparisons are displayed for clarity. MS, multi-system; CLAIM, Checklist for Artificial Intelligence in Medical Imaging; CI, confidence interval.

Table 4. Univariable and multivariable analysis of the study-level data to identify factors related to high and low CLAIM adherence

Variable	Category[1]	Univariable analysis				Multivariable analysis			
		CLAIM score		CLAIM compliance		CLAIM score		CLAIM compliance	
		Estimate	*P*	Estimate	*P*	Estimate	*P*	Estimate	*P*
Publication year	-	0.069	**0.028**	0.092	**0.010**	0.110	**0.007**	0.095	**0.028**
Radiology subfield	Dental	−2.590	**0.026**	−16.748	0.986	−2.672	**0.025**	−16.669	0.986
	Cardiovascular	14.585	0.977	16.384	0.985	15.722	0.984	16.998	0.982
	Genitourinary	0.272	0.760	1.070	0.221	0.542	0.639	0.643	0.474
	Neuro	−0.598	0.149	−0.418	0.265	−0.336	0.483	−0.025	0.953
	Chest	−1.613	**0.001**	0.274	0.548	−1.963	**<0.001**	−0.079	0.876
	Pediatric	−1.466	**0.014**	−1.345	**0.030**	−1.241	0.059	−1.106	0.091
	Musculoskeletal	−0.267	0.565	−0.155	0.713	−0.361	0.487	−0.075	0.869
Publication type	Print	0.387	0.700	1.014	0.382	-	-	-	-
	Preprint	−0.916	0.430	1.686	0.188	-	-	-	-
Scope of journals	Radiology or imaging-related	0.314	0.123	0.278	0.163	-	-	-	-
Focus of journals	AI-focused	0.256	0.652	−0.534	0.346	-	-	-	-
h5 index of journal	-	0.005	**0.013**	0.004	**0.027**	0.003	0.246	0.001	0.771
Impact factor quartile of journal	Q1	1.387	**<0.001**	0.750	**0.040**	1.754	**0.018**	2.577	**0.017**
	Q2	1.154	**0.004**	0.289	0.456	1.414	0.053	2.152	**0.046**
	Q3	0.386	0.454	−0.302	0.565	0.836	0.296	1.547	0.173
	Q4	0.128	0.836	−1.044	0.148	0.277	0.764	1.165	0.350

P values achieving statistical significance are in bold. CLAIM, Checklist for Artificial Intelligence in Medical Imaging; AI, artificial intelligence.
[1]Reference categories not shown.

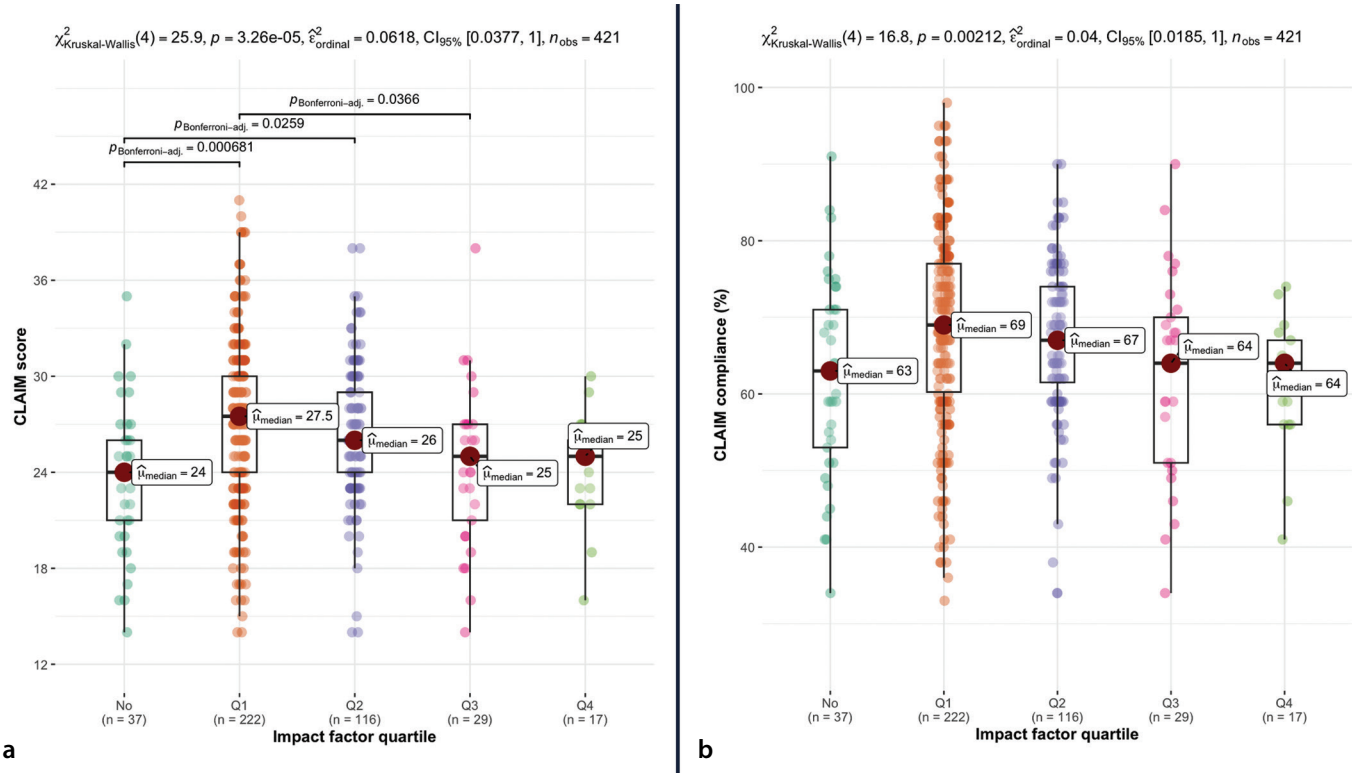

Figure 6. Box plots for the study-level analysis of the CLAIM score **(a)** and compliance **(b)** by impact factor quartile, with pairwise comparisons. The Kruskal–Wallis test showed statistically significant differences across all categories in both analyses **(a, b)**. Only statistically significant pairwise comparisons are displayed for clarity. CLAIM, Checklist for Artificial Intelligence in Medical Imaging; CI, confidence interval.

be influenced by the contribution of specific authors or research groups.

Eleven items were underreported in ≥50% of studies: de-identification methods (item#12), missing data handling (item#13), sample size determination (item#19), statistical significance and uncertainty (item#29), explainability methods (item#31), participant flow (item#33), demographic data (item#34), diagnostic accuracy estimates (item#36), failure analysis (item#37), registration details (item#40), and protocol access (item#41). This suggests challenges in fulfilling the CLAIM requirements, possibly due to inadequate knowledge, training, resource limitations, or the perceived irrelevance of certain items for specific study types. Interestingly, several of these items reflect broader challenges in AI research, such as securing adequate sample sizes, addressing uncertainty, enhancing model explainability to avoid the "black-box" problem, and promoting principles of open science, even if not explicitly stated. These 11 items, therefore, warrant particular attention when preparing AI manuscripts to improve the overall reporting transparency and rigor of AI research in medical imaging.

From the 35 eligible reviews, several key critiques were identified, including concerns about the inapplicability of certain items to all study types and the subjective nature of reporting decisions. Although the CLAIM 2024 update has addressed applicability by introducing three checklist options and leaving judgment to the evaluators,[10] subjective interpretation still remains a significant issue. Notably, our analysis revealed that CLAIM evaluations involved multiple readers in 85% of reviews, but only 11% assessed evaluation reliability, revealing a critical gap. Despite high reported reproducibility, such assessments need improved experimental settings to thoroughly investigate interpretation-related issues, as previously achieved for RQS.[57] Additionally, leveraging automated tools, such as those powered by large language models used for RQS,[58] might have the potential to help reduce subjectivity and improve consistency.

Based on the other critiques identified, future versions of the CLAIM can also be improved by simplifying definitions and improving clarity, removing subjective items based on reproducibility studies with rigorous analysis, and providing holistic guidance for interpreting manuscripts alongside their code. Additional improvements could include prioritizing items by assigning weights through evidence-based voting methods and developing user-friendly online tools, similar to the METhodological RadiomICs Score (METRICS),[59] for an adherence assessment that considers item applicability. These refinements would help streamline CLAIM evaluations and improve their utility for the medical imaging community.

Previous studies

To the best of our knowledge, no research has yet been conducted to evaluate CLAIM adherence by synthesizing data from both systematic and non-systematic reviews, providing a comprehensive overview of the topic. However, similar efforts have been made in the field of radiomics research,[23,60,61] particularly with the RQS,[62] which is widely regarded as the standard for assessing the methodological quality of radiomics studies, although recent alternatives have emerged.[59]

In 2023, Spadarella et al.[60], who first published their research online in 2022, conducted a review-level analysis of 44 reviews. They reported a median RQS of 21%. Later, in late 2024, Kocak et al.[23] deepened the analysis by performing a study-level analysis of 1,574 unique papers from 89 reviews, finding a median RQS of 31%. In 2025, in another very recent coincidental and independent study, Barry et al.[61] conducted a multi-lev-

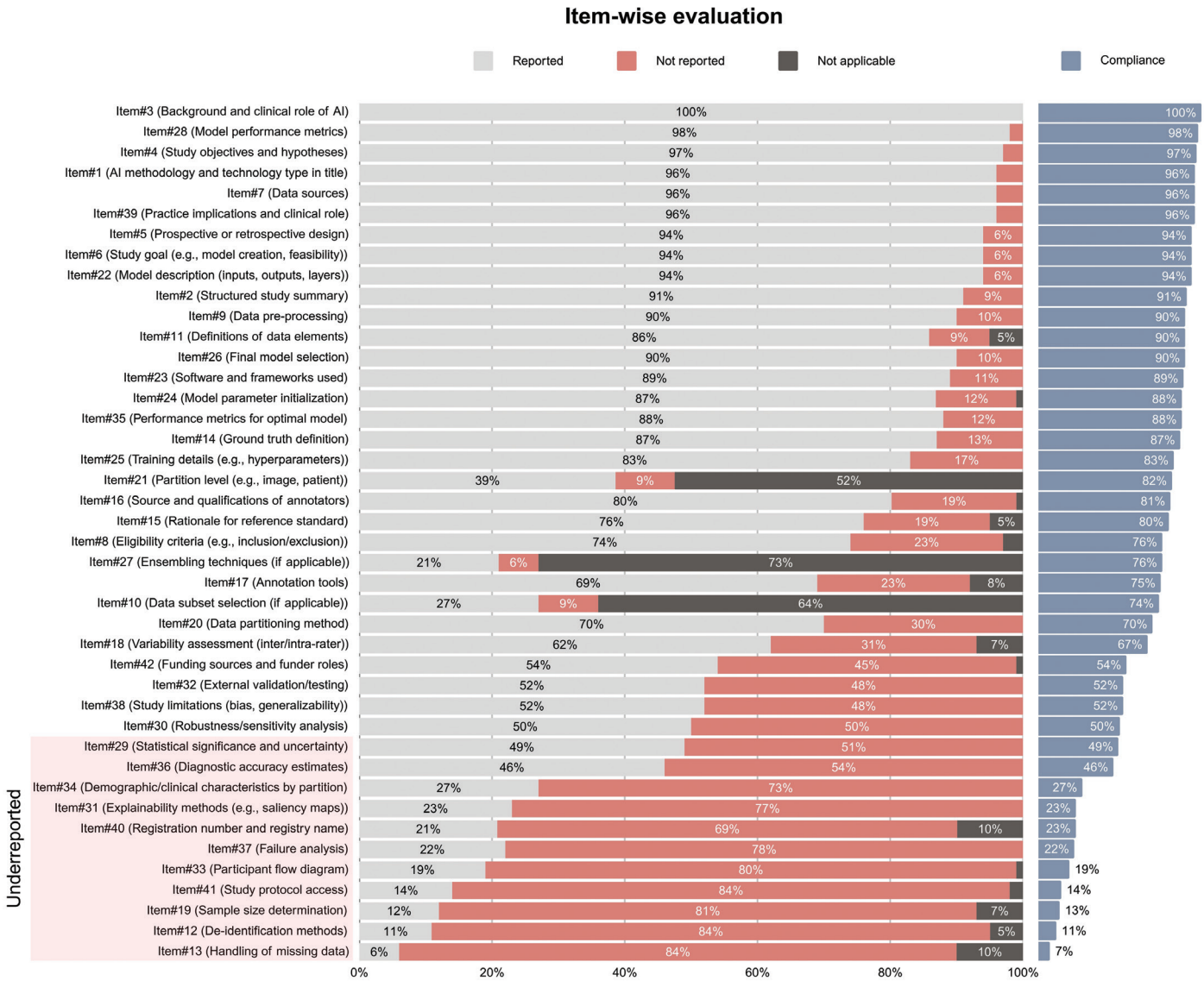

Figure 7. Item-wise analysis of the study-level data, ranked by compliance rates [calculated as follows: reported / (reported + not reported) × 100], considering the applicability of items. The compliance rates are based on the actual number of publications that reported or did not report each item. Note that item names have been abbreviated.

el meta-analysis of 3,258 RQS assessments from 130 systematic reviews as a continuation of the earlier study by Spadarella et al.[60], reporting an overall mean RQS of 9.4 ± 6.4 (95% confidence interval, 9.1–9.6) [26.1% ± 17.8% (25.3%–26.7%)]. It is important to note, however, that these RQS scores are not directly comparable to CLAIM adherence, as the two tools serve different purposes: RQS assesses the methodological quality of radiomics research, whereas the CLAIM focuses on reporting the quality of medical imaging AI research.

Furthermore, our results can be compared with those reported in the studies synthesized for this research.[12-22,27-50] In the review articles evaluated in the review-level analysis, the raw CLAIM scores ranged from 20 to 40, whereas the CLAIM adherence rates differed widely between 41% and 81%. This considerable variability underscores the inconsistent adherence to the CLAIM observed across the literature, highlighting the critical importance of our study in addressing these gaps.

Strengths and limitations

This study provides several strengths with notable implications for evaluating AI reporting quality in medical imaging. First, integrating data from multiple reviews offers a comprehensive assessment, unlike topic-specific studies, and provides a generalizable understanding of reporting practices. Second, our two-step analysis delivers both a broad overview and detailed insights, enabling item-wise evaluation to pinpoint areas needing particular improvement. Third, we identified factors associated with CLAIM adherence, offering actionable insights for enhancing reporting standards. Fourth, we presented two adherence metrics (the CLAIM score and compliance), facilitating comparability with other studies and setting a benchmark for future research. Finally, our analysis of critiques from eligible reviews offers valuable feedback to guide future updates to the CLAIM guidelines beyond 2024 and new alternative AI checklists.[10]

Our study has several limitations that should be carefully considered when inter-

Table 5. Item-wise correlation between reporting status and online publication year

CLAIM items[1]	Pre- and post-publication of CLAIM			Post-publication of CLAIM		
	rho	*P*	flag[2]	rho	*P*	flag[2]
Item#1 (AI methodology and technology type in title)	−0.097	0.046	*	−0.074	0.281	
Item#2 (Structured study summary)	0.034	0.491		0.022	0.748	
Item#3 (Background and clinical role of AI)	−0.038	0.435		0.071	0.300	
Item#4 (Study objectives and hypotheses)	−0.131	0.007	**	−0.162	0.018	*
Item#5 (Prospective or retrospective design)	0.092	0.060		0.017	0.806	
Item#6 (Study goal, e.g., model creation, feasibility)	−0.098	0.045	*	0.046	0.502	
Item#7 (Data sources)	0.024	0.626		−0.009	0.899	
Item#8 (Eligibility criteria, e.g., inclusion/exclusion)	−0.055	0.261		−0.014	0.842	
Item#9 (Data pre-processing)	−0.086	0.078		−0.217	0.001	**
Item#10 (Data subset selection, if applicable)	0.191	<0.001	***	0.225	<0.001	***
Item#11 (Definitions of data elements)	−0.220	<0.001	***	−0.057	0.405	
Item#12 (De-identification methods)	0.099	0.042	*	0.068	0.322	
Item#13 (Handling of missing data)	0.134	0.006	**	0.110	0.110	
Item#14 (Ground truth definition)	−0.057	0.240		−0.137	0.045	*
Item#15 (Rationale for reference standard)	−0.205	<0.001	***	−0.069	0.312	
Item#16 (Source and qualifications of annotators)	−0.078	0.111		−0.153	0.025	*
Item#17 (Annotation tools)	−0.244	<0.001	***	−0.092	0.180	
Item#18 [Variability assessment (inter/intra-rater)]	−0.211	<0.001	***	−0.121	0.078	
Item#19 (Sample size determination)	0.220	<0.001	***	0.250	<0.001	***
Item#20 (Data partitioning method)	0.140	0.004	**	−0.112	0.102	
Item#21 (Partition level, e.g., image, patient)	0.345	<0.001	***	0.116	0.091	
Item#22 [Model description (inputs, outputs, layers)]	0.036	0.462		−0.109	0.110	
Item#23 (Software and frameworks used)	−0.127	0.009	**	−0.134	0.050	
Item#24 (Model parameter initialization)	−0.124	0.011	*	−0.176	0.010	*
Item#25 (Training details, e.g., augmentation, hyperparameters)	0.141	0.004	**	−0.123	0.073	
Item#26 (Final model selection)	−0.057	0.246		−0.117	0.088	
Item#27 (Ensemble techniques, if applicable)	0.186	<0.001	***	0.167	0.014	*
Item#28 (Model performance metrics)	−0.076	0.119		−0.129	0.060	
Item#29 (Statistical significance and uncertainty)	0.026	0.594		0.060	0.386	
Item#30 (Robustness/sensitivity analysis)	0.022	0.656		0.015	0.831	
Item#31 (Explainability methods, e.g., saliency maps)	0.222	<0.001	***	−0.002	0.982	
Item#32 (External validation/testing)	0.009	0.846		0.098	0.152	
Item#33 (Participant flow diagram)	0.356	<0.001	***	0.202	0.003	**
Item#34 (Demographic/clinical characteristics by partition)	0.195	<0.001	***	0.126	0.065	
Item#35 (Performance metrics for optimal model)	−0.020	0.684		−0.097	0.158	
Item#36 (Diagnostic accuracy estimates)	0.101	0.039	*	0.172	0.012	*
Item#37 (Failure analysis)	0.075	0.125		−0.009	0.892	
Item#38 [Study limitations (bias, uncertainty, generalizability)]	0.173	<0.001	***	0.107	0.119	
Item#39 (Practice implications and clinical role)	−0.270	<0.001	***	−0.298	<0.001	***
Item#40 (Registration number and registry name)	−0.141	0.004	**	−0.093	0.174	
Item#41 (Study protocol access)	0.160	0.001	**	−0.075	0.275	
Item#42 (Funding sources and funder roles)	0.207	<0.001	***	0.092	0.178	

[1] Note that item names have been abbreviated; [2]* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$. CLAIM, Checklist for Artificial Intelligence in Medical Imaging.

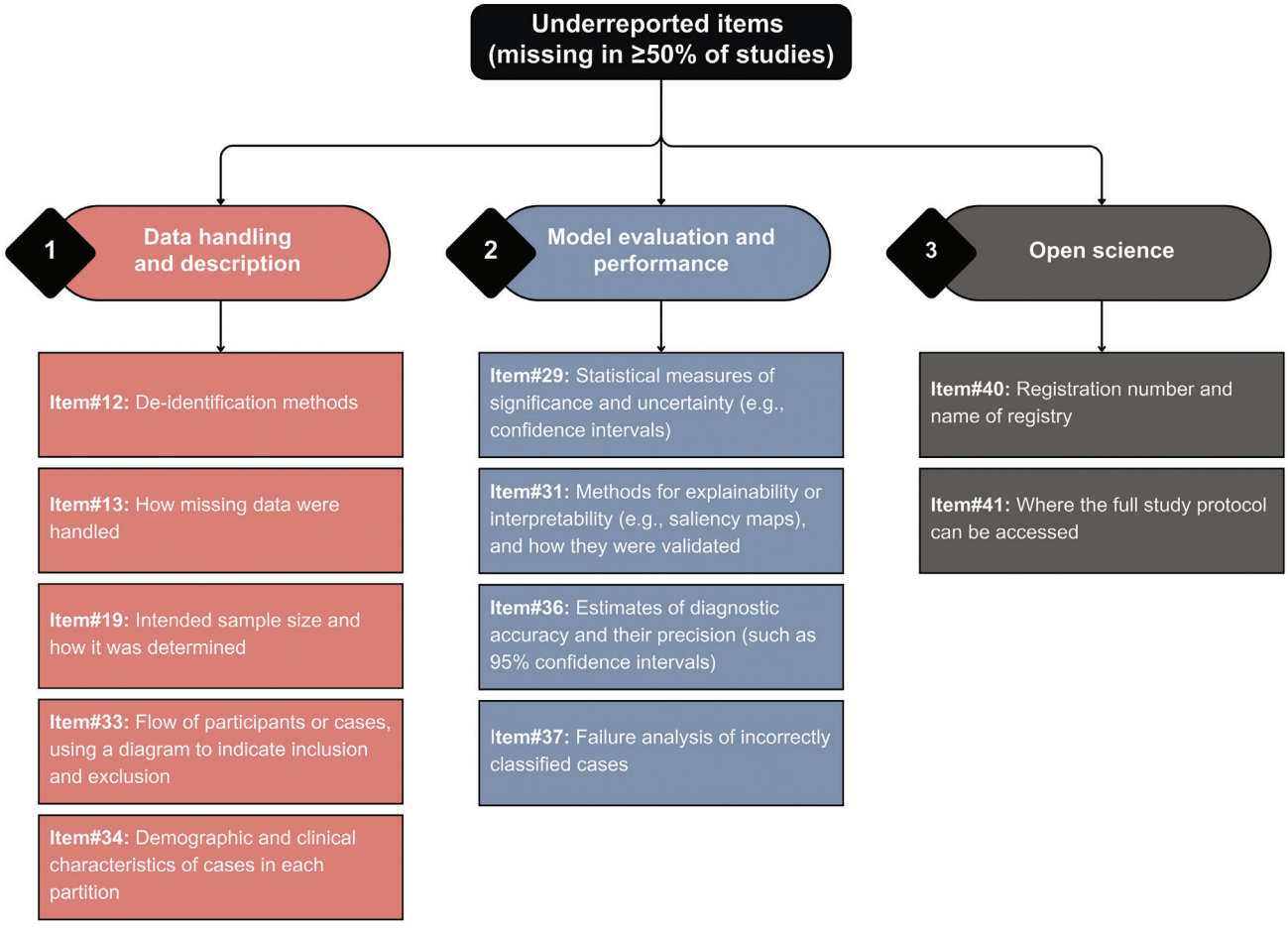

Figure 8. Eleven underreported items (i.e., missing in ≥50% of studies), categorized by relevant domains.

Table 6. Critiques identified in the analysis of the 35 review papers eligible for review- or study-level analyses

Category	Critique identified about the CLAIM with representative source articles
Fulfillment	Certain items may be viewed as overly strict or difficult to meet[43]
	Certain items are too technical, requiring advanced engineering or statistical knowledge[14]
Applicability	Some items are not applicable to all study types[12-14,30,39]
Feasibility and practicality	Some items may be impractical or infeasible in real-world settings[22]
Structure	Dividing the checklist into distinct sections sometimes complicates quality assessment[39]
Interpretation	Deciding if an item is sufficiently reported is subjective[13,39,44]
	Certain items may be viewed as vague or lack clarity in their current form[22]
	Certain items provide limited guidance on holistically interpreting a manuscript alongside its code[45]
Relative importance	Certain items may be more crucial than others but are currently weighted equally[13,39]
Scoring	Lack of standardized score or compliance calculation strategy[44]

CLAIM, Checklist for Artificial Intelligence in Medical Imaging.

preting the results. First, this study was not registered (e.g., in PROSPERO). This decision was due to the unique nature of conducting a collective review of previous reviews of the CLAIM. Given the limited number of studies employing a similar strategy, and despite our group's experience with other guidelines, the methodology required adaptations based on the challenges and limitations encountered during data collection and analysis. These evolving methodological adjustments made it difficult to provide a fully transparent outline of the approach at the outset. Second, this research was limited to three databases, PubMed, Scopus, and Google Scholar, which we selected based on their broad coverage and relevance to the field, according to our experience. However, we acknowledge that the inclusion of additional databases, such as Embase and Web of Science, could further improve the comprehensiveness of the search. Third, the assessment of reporting quality was based solely on the CLAIM (2020 version). In the future, other AI-specific reporting guidelines, such as CONSORT-AI and TRIPOD-AI, could be considered to provide a more comprehensive evaluation of reporting standards.[63] Fourth, many articles were published before the CLAIM guidelines were introduced in 2020. However, the goal of this study was to highlight the overall state of reporting quality in the field, with some analyses covering both pre- and post-guideline periods. Fifth, our analysis focused solely on reporting quality and did not include evaluating the studies' actual impact, such as citation counts; there may not yet have been sufficient time for recent studies to have accumulated citations for meaningful comparisons. Additionally, the scope of our study is limited to exploring other factors that could affect the clinical translation of AI, such as methodological quality. Evaluating these factors may require supplementary tools, such as METRICS.[59] Sixth, this study was conducted after the CLAIM 2024 update.[10] Although the main framework of the original CLAIM was preserved,[8] earlier findings might have better informed the current update but could still aid future revisions and new guidelines. Seventh, the results of this study rely on prior systematic and non-systematic reviews as well as the expertise of the evaluators involved in those studies. The potential limited familiarity with certain aspects of the CLAIM in those articles and inconsistencies may have influenced the findings of this study. Eighth, due to the lack of a standard checkbox format in the initial CLAIM, consideration of item applicability may vary among reviews, potentially influencing adherence results, although both the CLAIM score and CLAIM compliance were assessed in the two-level analysis. Ninth, extracting data from systematic reviews can be subjective and may vary depending on the readers' experience. To minimize potential errors, we implemented a rigorous process involving the cross-checking of extracted data and resolving disagreements through consensus or by consulting an experienced reader, when necessary, at different stages of the study. Finally, the number of studies included in the study-level analysis was smaller than the number of studies represented in the review articles analyzed at the review level. However, to gain item-level insights, it was essential to conduct the analysis at the individual study level, as this granularity could not have been achieved at the review level. The sample size for the study-level analysis was determined merely by the availability of data in the existing literature, which may have introduced some degree of bias. Therefore, the findings should be interpreted with this limitation in mind.

In conclusion, this study provides a comprehensive evaluation of CLAIM adherence in the medical imaging AI literature, revealing significant variability and highlighting areas for improvement. Our two-level analysis, encompassing review- and study-level data, identified substantial reporting gaps, with a third of checklist items often omitted. Factors such as publication year, journal impact quartiles, and subfield-specific differences emerged as key independent predictors of adherence, underscoring the role of high-impact journals and tailored strategies for different subfields. The CLAIM compliance rate was highlighted as a more objective and fairer metric for adherence assessment. Additionally, several important critiques of the CLAIM were identified, providing valuable insights for researchers and developers. We hope these findings serve as actionable guidance for the scientific community to enhance transparency, reproducibility, and reporting quality in AI studies.

Acknowledgements

The authors wish to express their sincere gratitude to the anonymous reviewers for their insightful comments and constructive feedback, which have significantly improved the quality of this paper.

The language of this manuscript was checked and improved by ChatGPT (4o). This tool was used solely to improve the clarity and quality of the content originally written by the authors. The authors conducted strict supervision after using this tool.

Footnotes

Conflict of Interest

Burak Koçak, MD, is Section Editor in Diagnostic and Interventional Radiology. He had no involvement in the peer-review of this article and had no access to information regarding its peer-review. Other authors have nothing to disclose.

References

1. Kocak B, Baessler B, Cuocolo R, Mercaldo N, Pinto Dos Santos D. Trends and statistics of artificial intelligence and radiomics research in radiology, nuclear medicine, and medical imaging: bibliometric analysis. *Eur Radiol*. 2023;33(11):7542-7555. [CrossRef]
2. Nensa F, Pinto Dos Santos D, Dietzel M. Beyond accuracy: reproducibility must lead AI advances in radiology. *Eur J Radiol*. 2024;180:111703. [CrossRef]
3. Beam AL, Manrai AK, Ghassemi M. Challenges to the reproducibility of machine learning models in health care. *JAMA*. 2020;323(4):305-306. [CrossRef]
4. Klement W, El Emam K. Consolidated reporting guidelines for prognostic and diagnostic machine learning modeling studies: development and validation. *J Med Internet Res*. 2023;25:e48763. [CrossRef]
5. Vasey B, Novak A, Ather S, Ibrahim M, McCulloch P. DECIDE-AI: a new reporting guideline and its relevance to artificial intelligence studies in radiology. *Clin Radiol*. 2023;78(2):130-136. [CrossRef]
6. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26(9):1364-1374. [CrossRef]
7. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. [CrossRef]
8. 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. [CrossRef]
9. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*. 2015;351:h5527. [CrossRef]
10. Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for Artificial Intelligence in Medical

Imaging (CLAIM): 2024 update. *Radiol Artif Intell*. 2024;6(4):e240300. [CrossRef]

11. Pandis N, Fedorowicz Z. The international EQUATOR network: enhancing the quality and transparency of health care research. *J Appl Oral Sci*. 2011;19(5):0. [CrossRef]
12. Belue MJ, Harmon SA, Lay NS, 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(2):134-145. [CrossRef]
13. 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(5):907-913. [CrossRef]
14. Si L, Zhong J, Huo J, et al. Deep learning in knee imaging: a systematic review utilizing a Checklist for Artificial Intelligence in Medical Imaging (CLAIM). *Eur Radiol*. 2022;32(2):1353-1361. [CrossRef]
15. Abdulaal L, Maiter A, Salehi M, et al. A systematic review of artificial intelligence tools for chronic pulmonary embolism on CT pulmonary angiography. *Front Radiol*. 2024;4:1335349. [CrossRef]
16. Alipour E, Pooyan A, Shomal Zadeh F, Darbandi AD, Bonaffini PA, Chalian M. Current status and future of artificial intelligence in MM imaging: a systematic review. *Diagnostics (Basel)*. 2023;13(21):3372. [CrossRef]
17. Sushentsev N, Moreira Da Silva N, Yeung M, et al. Comparative performance of fully-automated and semi-automated artificial intelligence methods for the detection of clinically significant prostate cancer on MRI: a systematic review. *Insights Imaging*. 2022;13(1):59. [CrossRef]
18. Ozkara BB, Chen MM, Federau C, et al. Deep Learning for Detecting Brain Metastases on MRI: a systematic review and meta-analysis. *Cancers*. 2023;15(2):334. [CrossRef]
19. Alabed S, Maiter A, Salehi M, 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. [CrossRef]
20. Karabacak M, Ozkara BB, Ozturk A, et al. Radiomics-based machine learning models for prediction of medulloblastoma subgroups: a systematic review and meta-analysis of the diagnostic test performance. *Acta Radiol*. 2023;64(5):1994-2003. [CrossRef]
21. 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(1):vdac081. [CrossRef]
22. Kim DY, Oh HW, Suh CH. 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. *Korean J Radiol*. 2023;24(12):1179-1189. [CrossRef]
23. Kocak B, Keles A, Kose F, Sendur A. Quality of radiomics research: comprehensive analysis of 1574 unique publications from 89 reviews. *Eur Radiol*. 2024. [CrossRef]
24. Briscoe S, Abbott R, Lawal H, Shaw L, Coon JT. Feasibility and desirability of screening search results from Google Search exhaustively for systematic reviews: a cross-case analysis. *Res Synth Methods*. 2023;14(3):427-437. [CrossRef]
25. Luo D, Wan X, Liu J, Tong T. Optimally estimating the sample mean from the sample size, median, mid-range, and/or mid-quartile range. *Stat Methods Med Res*. 2018;27(6):1785-1805. [CrossRef]
26. Wan X, Wang W, Liu J, Tong T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC Med Res Methodol*. 2014;14(1):135. [CrossRef]
27. Assadi H, Alabed S, Maiter A, et al. The role of artificial intelligence in predicting outcomes by cardiovascular magnetic resonance: a comprehensive systematic review. *Medicina (Kaunas)*. 2022;58(8):1087. [CrossRef]
28. Bedrikovetski S, Seow W, Kroon HM, Traeger L, Moore JW, Sammour T. Artificial intelligence for body composition and sarcopenia evaluation on computed tomography: a systematic review and meta-analysis. *Eur J Radiol*. 2022;149:110218. [CrossRef]
29. Belue MJ, Turkbey B. Tasks for artificial intelligence in prostate MRI. *Eur Radiol Exp*. 2022;6(1):33. [CrossRef]
30. Bleker J, Kwee TC, Yakar D. Quality of multicenter studies using MRI radiomics for diagnosing clinically significant prostate cancer: a systematic review. *Life (Basel)*. 2022;12(7):946. [CrossRef]
31. Cerdá-Alberich L, Solana J, Mallol P, et al. MAIC-10 brief quality checklist for publications using artificial intelligence and medical images. *Insights Imaging*. 2023;14(1):11. [CrossRef]
32. Dagher R, Ozkara BB, Karabacak M, et al. Artificial intelligence/machine learning for neuroimaging to predict hemorrhagic transformation: Systematic review/meta-analysis. *J Neuroimaging*. 2024;34(5):505-514. [CrossRef]
33. Hardacre CJ, Robertshaw JA, Barratt SL, et al. Diagnostic test accuracy of artificial intelligence analysis of cross-sectional imaging in pulmonary hypertension: a systematic literature review. *Br J Radiol*. 2021;94(1128):20210332. [CrossRef]
34. Hickman SE, Woitek R, Le EPV, et al. Machine learning for workflow applications in screening mammography: systematic review and meta-analysis. *Radiology*. 2022;302(1):88-104. [CrossRef]
35. Hu J, Wang Y, Guo D, et al. Diagnostic performance of magnetic resonance imaging-based machine learning in Alzheimer's disease detection: a meta-analysis. *Neuroradiology*. 2023;65(3):513-527. [CrossRef]
36. Hwang EJ, Jeong WG, David PM, Arentz M, Ruhwald M, Yoon SH. AI for detection of tuberculosis: implications for global health. *Radiol Artif Intell*. 2024;6(2):e230327. [CrossRef]
37. Jia LL, Zhao JX, Pan NN, et al. Artificial intelligence model on chest imaging to diagnose COVID-19 and other pneumonias: a systematic review and meta-analysis. *Eur J Radiol Open*. 2022;9:100438. [CrossRef]
38. Karabacak M, Ozkara BB, Mordag S, Bisdas S. Deep learning for prediction of isocitrate dehydrogenase mutation in gliomas: a critical approach, systematic review and meta-analysis of the diagnostic test performance using a Bayesian approach. *Quant Imaging Med Surg*. 2022;12(8):4033-4046. [CrossRef]
39. Lans A, Pierik RJB, Bales JR, et al. Quality assessment of machine learning models for diagnostic imaging in orthopaedics: A systematic review. *Artif Intell Med*. 2022;132:102396. [CrossRef]
40. Le VNT, Kim JG, Yang YM, Lee DW. Evaluating the Checklist for Artificial Intelligence in Medical Imaging (CLAIM)-based quality of reports using convolutional neural network for odontogenic cyst and tumor detection. *Appl Sci*. 2021;11(20):9688. [CrossRef]
41. O'Shea RJ, Sharkey AR, Cook GJR, Goh V. Systematic review of research design and reporting of imaging studies applying convolutional neural networks for radiological cancer diagnosis. *Eur Radiol*. 2021;31(10):7969-7983. [CrossRef]
42. Raj M, Ayub A, Pal AK, et al. Diagnostic accuracy of artificial intelligence-based algorithms in automated detection of neck of femur fracture on a plain radiograph: a systematic review and meta-analysis. *Indian J Orthop*. 2024;58(5):457-469. [CrossRef]
43. Roberts M, Driggs D, Thorpe M, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell*. 2021;3(3):199-217. [CrossRef]
44. Roest C, Fransen SJ, Kwee TC, Yakar D. Comparative performance of deep learning and radiologists for the diagnosis and localization of clinically significant prostate cancer at MRI: a systematic review. *Life (Basel)*. 2022;12(10):1490. [CrossRef]
45. Sivanesan U, Wu K, McInnes MDF, Dhindsa K, Salehi F, van der Pol CB. Checklist for Artificial Intelligence in Medical Imaging Reporting adherence in peer-reviewed and preprint manuscripts with the highest altmetric attention scores: a meta-research study. *Can Assoc Radiol J*. 2023;74(2):334-342. [CrossRef]

46. 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(10):1127-1147. [CrossRef]
47. Wang TW, Hong JS, Huang JW, Liao CY, Lu CF, Wu YT. Systematic review and meta-analysis of deep learning applications in computed tomography lung cancer segmentation. *Radiother Oncol*. 2024;197:110344. [CrossRef]
48. Wang TW, Hsu MS, Lee WK, et al. Brain metastasis tumor segmentation and detection using deep learning algorithms: a systematic review and meta-analysis. *Radiother Oncol*. 2024;190:110007. [CrossRef]
49. Zhong J, Hu Y, Zhang G, 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(1):138. [CrossRef]
50. Zhong J, Xing Y, Zhang G, et al. A systematic review of radiomics in giant cell tumor of bone (GCTB): the potential of analysis on individual radiomics feature for identifying genuine promising imaging biomarkers. *J Orthop Surg Res*. 2023;18(1):414. [CrossRef]
51. Agha RA, Fowler AJ, Limb C, et al. Impact of the mandatory implementation of reporting guidelines on reporting quality in a surgical journal: a before and after study. *Int J Surg*. 2016;30:169-172. [CrossRef]
52. Koçak B, Keleş A, Köse F. Meta-research on reporting guidelines for artificial intelligence: are authors and reviewers encouraged enough in radiology, nuclear medicine, and medical imaging journals? *Diagn Interv Radiol*. 2024;30(5):291-298. [CrossRef]
53. Zhong J, Xing Y, Lu J, et al. The endorsement of general and artificial intelligence reporting guidelines in radiological journals: a meta-research study. *BMC Med Res Methodol*. 2023;23(1):292. [CrossRef]
54. Kocak B, Keles A, Akinci D'Antonoli T. Self-reporting with checklists in artificial intelligence research on medical imaging: a systematic review based on citations of CLAIM. *Eur Radiol*. 2024;34(4):2805-2815. [CrossRef]
55. Kocak B, Ponsiglione A, Stanzione A, et al. CLEAR guideline for radiomics: early insights into current reporting practices endorsed by EuSoMII. *Eur J Radiol*. 2024;181:111788. [CrossRef]
56. Korevaar DA, Wang J, van Enst WA, et al. Reporting diagnostic accuracy studies: some improvements after 10 years of STARD. *Radiology*. 2015;274(3):781-789. [CrossRef]
57. Akinci D'Antonoli T, Cavallo AU, Vernuccio F, et al. Reproducibility of radiomics quality score: an intra- and inter-rater reliability study. *Eur Radiol*. 2024;34(4):2791-2804. [CrossRef]
58. Mese I, Kocak B. ChatGPT as an effective tool for quality evaluation of radiomics research. *Eur Radiol*. 2024. [CrossRef]
59. Kocak B, Akinci D'Antonoli T, Mercaldo N, et al. METhodological RadiomICs Score (METRICS): a quality scoring tool for radiomics research endorsed by EuSoMII. *Insights Imaging*. 2024;15(1):8. [CrossRef]
60. Spadarella G, Stanzione A, Akinci D'Antonoli T, et al. Systematic review of the radiomics quality score applications: an EuSoMII Radiomics Auditing Group Initiative. *Eur Radiol*. 2023;33(3):1884-1894. [CrossRef]
61. Barry N, Kendrick J, Molin K, et al. Evaluating the impact of the Radiomics Quality Score: a systematic review and meta-analysis. *Eur Radiol*. 2025. [CrossRef]
62. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*. 2017;14(12):749-762. [CrossRef]
63. Park SH, Suh CH. Reporting Guidelines for Artificial Intelligence Studies in Healthcare (for Both Conventional and Large Language Models): what's new in 2024. *Korean J Radiol*. 2024;25(8):687-690. [CrossRef]