Original Article
Yonsei Med J 2024 Oct;65(10):611-618
https://doi.org/10.3349/ymj.2024.0007
Yonsei Medical Journal
YMJ
pISSN: 0513-5796 · eISSN: 1976-2437

Need for Transparency and Clinical Interpretability in Hemorrhagic Stroke Artificial Intelligence Research: Promoting Effective Clinical Application

Chae Young Lim, Beomseok Sohn, Minjung Seong, Eung Yeop Kim, Sung Tae Kim, and So Yeon Won
Department of Radiology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea.

Purpose: This study aimed to evaluate the quality of artificial intelligence (AI)/machine learning (ML) studies on hemorrhagic stroke using the Minimum Information for Medical AI Reporting (MINIMAR) and Minimum Information About Clinical Artificial Intelligence Modeling (MI-CLAIM) frameworks to promote clinical application.
Materials and Methods: PubMed, MEDLINE, and Embase were searched for AI/ML studies on hemorrhagic stroke. Out of the 531 articles found, 29 relevant original research articles were included. MINIMAR and MI-CLAIM scores were assigned by two experienced radiologists to assess the quality of the studies.
Results: We analyzed 29 investigations that utilized AI/ML in the field of hemorrhagic stroke, involving a median of 224.5 patients. The majority of studies focused on diagnostic outcomes using computed tomography scans (89.7%) and were published in computer science journals (48.3%). The overall adherence rates to reporting guidelines, as assessed through the MINIMAR and MI-CLAIM frameworks, were 47.6% and 46.0%, respectively. In MINIMAR, none of the studies reported the socioeconomic status of the patients or how missing values had been addressed. In MI-CLAIM, only two studies applied model-examination techniques to improve model interpretability. Transparency and reproducibility were limited, as only 10.3% of the studies had publicly shared their code. Cohen's kappa between the two radiologists was 0.811 and 0.779 for MINIMAR and MI-CLAIM, respectively.
Conclusion: The overall reporting quality of published AI/ML studies on hemorrhagic stroke is suboptimal. It is necessary to incorporate model examination techniques for interpretability and promote code openness to enhance transparency and increase the clinical applicability of AI/ML studies.

Key Words: Hemorrhagic stroke, artificial intelligence, machine learning, reporting guidelines

INTRODUCTION

In the field of stroke, hemorrhagic stroke is a significant concern, accounting for approximately 10%–15% of all strokes and exhibiting high rates of morbidity and mortality.[1,2] In 2016, the global age-standardized incidence rate for intracerebral hemorrhage was 22.2 per 100000 person-years, with hemorrhagic stroke associated with a 30-day mortality rate of up to 40%.[3,4] Furthermore, survivors often experience severe long-term disabilities.[5] Early detection of hemorrhagic stroke is crucial for improving patient outcomes, as timely intervention can significantly reduce morbidity and mortality.[6,7] Additionally, the prediction of hematoma expansion or other complications, such as herniation or hydrocephalus, as well as prognostication of neurological outcomes, hold great importance. As a result, the field has seen a rise in artificial intelligence (AI)/machine learning (ML)-based publications to fulfill the requirements of timely diagnosis, prediction, and prognostication.

AI research has been conducted in various fields using diverse methods, leading to a lack of uniformity. This lack of consistency hampers the translation of AI research into clinical

Received: March 6, 2024 **Revised:** May 8, 2024
Accepted: May 20, 2024 **Published online:** July 18, 2024
Corresponding author: So Yeon Won, MD, PhD, Department of Radiology, Samsung Medical Center, Sungkyunkwan University School of Medicine, 81 Irwon-ro, Gangnam-gu, Seoul 06351, Korea.
E-mail: wsy0622@naver.com

•The authors have no potential conflicts of interest to disclose.

© Copyright: Yonsei University College of Medicine 2024
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

application. Consequently, there is a need for consistent evaluation criteria and guidelines. Recently, a few checklists and quality assessment tools have been developed and are in development specifically for AI research, such as Minimum Information for Medical AI Reporting (MINIMAR) and Minimum Information About Clinical Artificial Intelligence Modeling (MI-CLAIM). These tools aim to assist researchers and reviewers in assessing the methodological rigor and reporting quality of AI/ML studies.[8-16] However, systematic evaluations of AI/ML papers on neuroimaging using these tools remain limited.

The objective of the present study was to evaluate the quality of ML and AI papers in the field of hemorrhagic stroke using these innovative tools. The study aimed to identify areas that need improvement and further development to enhance clinical application.

MATERIALS AND METHODS

Systematic search strategy and study selection

A comprehensive review was performed, including all clinical radiological papers that utilized AI/ML to address tasks related to hemorrhagic stroke. This study specifically included hospital-based human studies that employed these techniques to assist in the radiological diagnosis or intervention of patients. Studies utilizing radiography, computed tomography (CT), and magnetic resonance imaging (MRI) were considered. PubMed, MEDLINE (n=1166), and Embase (n=962) databases were searched on March 10, 2023 to collect all original research papers utilizing ML-based analysis published until March 10, 2023. Due to the characteristics of the current study, the authors did not have access to information that could identify individual participants. The terms used for the search are listed in Supplementary Material 1 (only online). Out of the 2128 papers identified in the search, 697 duplicate articles were removed, and an additional 900 articles were excluded for the following reasons: conference abstracts (n=411), not in the field of interest (n=390), review articles (n=73), non-AI studies (n=9), Erratum (n=9), non-human study (n=6), and short note (n=2). Among the remaining 531 articles, studies on ischemic stroke (n=416) and carotid artery disease (n=86) were also excluded. Finally, a total of 29 articles were included in the analysis (Supplementary Table 1, only online). The flow of study selection is depicted in Fig. 1.

AI reporting quality based on MINIMAR

Each article was assessed for the presence of the four essential components outlined in MINIMAR, which consist of 21 features. The four essential components are study population and setting, patient demographics, model architecture, and model evaluation. The MINIMAR evaluation was carried out by the reviewers, who received education on the MINIMAR system through a research conference prior to the assessment. Two reviewers (with 14 and 4 years of radiology experience, respectively) independently rated each paper using MINIMAR, focusing on the four components. Any disagreements between the reviewers were resolved through consensus. The MINIMAR checklist can be found in Supplementary Material 2 (only online).

AI reporting quality based on MI-CLAIM

The evaluation of each article involved assessing the four essential components of MI-CLAIM, which encompassed 18 features. The four essential components are study design, data and optimization, model performance, and model examination. Furthermore, the analysis of MI-CLAIM also included the task of categorizing into one of the four categories based on the

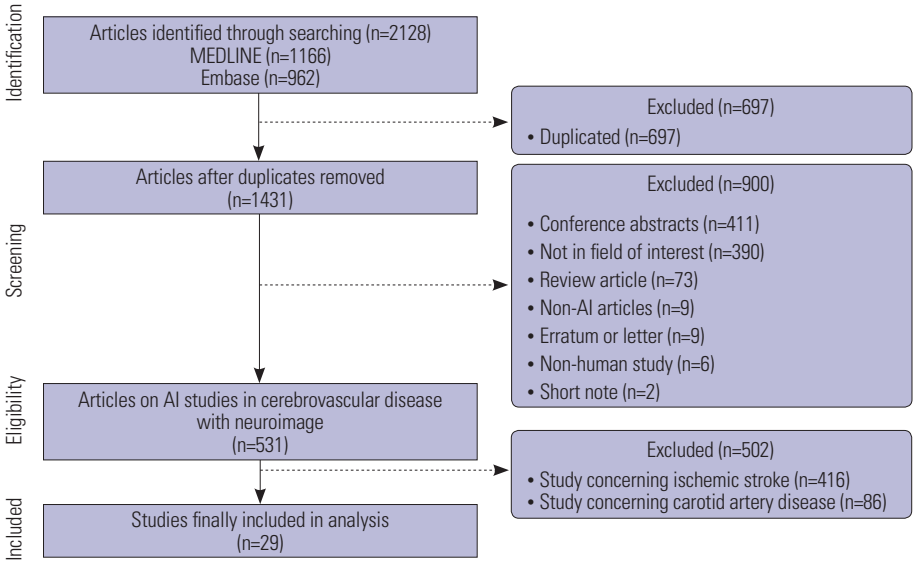

Fig. 1. Flowchart of study selection. AI, artificial intelligence.

https://doi.org/10.3349/ymj.2024.0007

level of transparency. The MI-CLAIM evaluation was conducted by the same reviewers, who received education on the MI-CLAIM system through a research conference prior to the assessment. Two reviewers (with 14 and 4 years of radiology experience, respectively) independently rated each paper using MI-CLAIM. The MI-CLAIM checklist is available in Supplementary Material 3 (only online).

Statistical analysis

The current study examined the quality of 29 articles, assessing 21 MINIMAR criteria and 18 MI-CLAIM criteria. To represent each article's compliance with the MINIMAR or MI-CLAIM criteria, the adherence rate was defined as the proportion of criteria met out of the total applicable criteria for each study, expressed as percentage. Cohen's kappa was calculated to estimate the interobserver agreement between the two reviewers regarding MINIMAR and MI-CLAIM compliance. Statistical analyses were conducted using R (version 4.0.2; R Foundation for Statistical Computing, Vienna, Austria).

RESULTS

Characteristics of the AI studies on hemorrhagic stroke

Fig. 2 and Supplementary Table 2 (only online) present the characteristics of the 29 AI/ML investigations included in this study. The median number of patients included in 28 out of 29 studies was 224.5 (range: 16–5244734). One study did not report the number of patients included. The publications included in the analysis consisted of 14 computer science journals (48.3%), 10 clinical journals (34.5%), and 5 radiology journals (17.2%). The outcomes of the AI/ML-based investigations were 19 diagnostic (65.5%), 7 predictive (24.1%), 2 efficiency (6.9%), and 1 combination of diagnostic and predictive (3.4%). Among the 29 studies, 26 (89.7%) utilized CT, 2 (6.9%) utilized MRI, and 1 (3.4%) utilized both CT and MRI. Three (10.3%) out of the 29 articles had publicly shared their codes.[17-19]

Adherence to reporting AI/ML research using MINIMAR

All 29 studies included in the analysis reported the task of their model, with 28 of them providing information on the output of

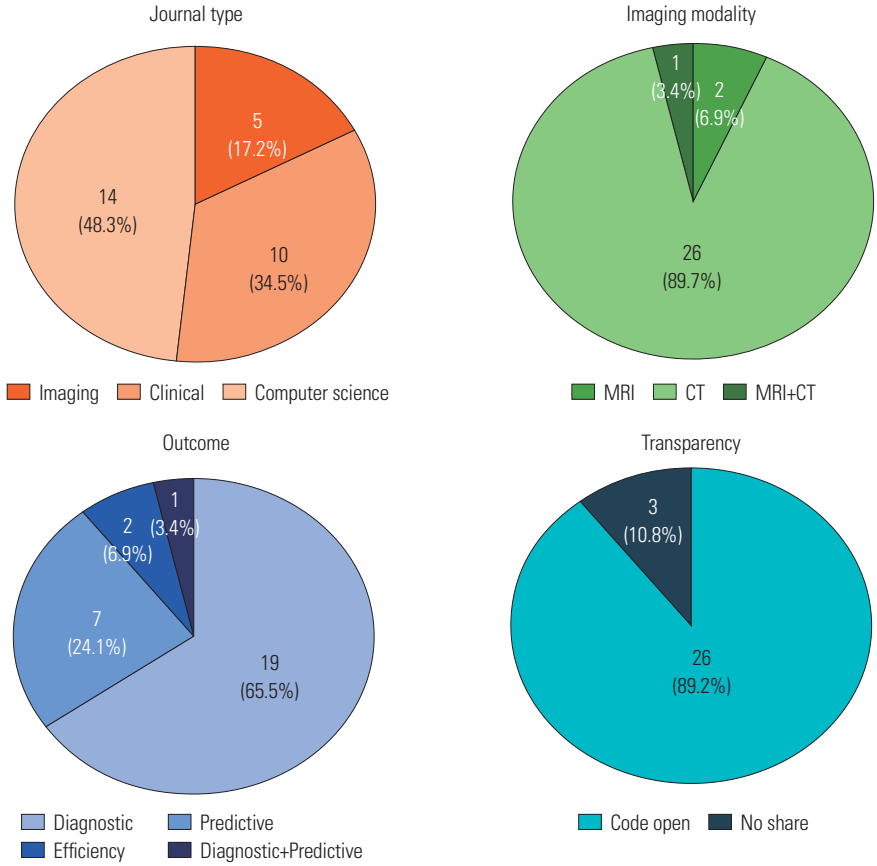

Fig. 2. Characteristics of AI/ML studies on hemorrhagic stroke. (A) The type of journal publication included 14 computer science journals (48.3%), 10 clinical journals (34.5%), and 5 imaging journals (17.2%). (B) Of the 29 studies, 26 (89.7%) utilized CT, 2 (6.9%) utilized MRI, and 1 (3.4%) utilized both CT and MRI. (C). Outcomes of AI/ML-based investigations were 19 diagnostic (65.5%), 7 predictive (24.1%), 2 efficiency (6.9%), and 1 combination of diagnostic and predictive (3.4%). (D) Three (10.3%) of 29 articles had shared their code publicly. AI, artificial intelligence; ML, machine learning.

their model. The initial evaluation between the two radiologists showed a Cohen's kappa of 0.811. After reaching a consensus, the overall adherence rate to MINIMAR was found to be 47.6%. Table 1 illustrates the conformity rates for the four essential components of MINIMAR: study population and setting, patient demographics, model architecture, and model evaluation. The adherence rates for each component were 77.6%, 25.5%, 55.6%, and 37.9%, respectively. The adherence to each MINIMAR item is also presented in Table 1. It was observed that none of the studies reported the socioeconomic status of the patients or how missing data had been addressed.

Adherence to reporting AI/ML research using MI-CLAIM

Cohen's kappa was 0.779 at the initial evaluation between the two radiologists. After consensus had been reached, the total MI-CLAIM adherence rate was 46.0%. Table 2 describes the fundamental conformance rate to reporting adherence to MI-CLAIM by its four essential components: study design, data and optimization, model performance, and model examination. The adherence rates of the components were 64.1%, 55.2%, 51.7%, and 20.7%, respectively. Table 2 shows the adherence to each MI-CLAIM item. Only two studies applied model examination techniques to improve model interpretability.[20,21] Only one study discussed the feasibility and interpretability of its suggested model.[22] Among the 29 papers, 26 (89.7%) excluding three[17-19] received the lowest category rating in terms of transparency level.

In our evaluation of the 29 articles included in the study, vari-

Table 1. Adherence to MINIMAR Item in 29 Included Hemorrhagic Stroke Studies

Total (21 items)	Value
Study population and setting	77.6%
Population	23 (79.3)
Setting	20 (69.0)
Data source	27 (93.1)
Cohort selection	20 (69.0)
Patient demographics	25.5%
Age	17 (58.6)
Sex	17 (58.6)
Race	2 (6.9)
Ethnicity	1 (3.4)
Socioeconomic status	0 (0)
Model architecture	55.6%
Model output	27 (93.1)
Target user	4 (13.8)
Data splitting	16 (55.2)
Gold standard	25 (86.2)
Model task	29 (100)
Model architecture	20 (69.0)
Features	8 (27.6)
Missingness	0 (0)
Model evaluation	37.9%
Optimization	10 (34.5)
Internal validation	19 (65.5)
External validation	12 (41.4)
Transparency	3 (10.3)
Mean adherence rate	47.6%

MINIMAR, Minimum Information for Medical AI Reporting.
Data are presented as n (%). Each number and its corresponding percentage in the criteria and category represent the count and proportion of articles that meet each specific criterion, respectively.

Table 2. Adherence to MI-CLAIM Items in 29 Included Hemorrhagic Stroke Studies

Total (18 items)	Value
Study design	64.1%
Clinical problem	26 (89.7)
Research question	28 (96.6)
Cohort characteristics	21 (72.4)
Cohort representing real-world	13 (44.8)
State of the art as comparison	5 (17.2)
Data and optimization	55.2%
Data origin and format	27 (93.1)
Data transformations	18 (62.1)
Test set independence	17 (58.6)
Model evaluation and best selection	2 (6.9)
Input*	
Structured	2 (6.9)
Unstructured	27 (93.1)
Model performance	51.7%
Performance metrics	23 (79.3)
Clinical utility metrics	17 (58.6)
Comparison with baseline models	5 (17.2)
Model examination	20.7%
Model examination technique	4 (13.8)
Examination technique 1-for structured e.g. SHAP†	1 (3.4)
Examination technique 2-for unstructured e.g. Grad-CAM†	3 (10.3)
Relevance discussion	16 (55.2)
Feasibility and interpretability discussion	1 (3.4)
Robustness discussion	3 (10.3)
Transparency	10.3%
Complete sharing of the code	3 (10.3)
Third party evaluation of the code	0 (0)
Release of a virtual machine	0 (0)
Mean adherence rate	46.0%

MI-CLAIM, Minimum Information About Clinical Artificial Intelligence Modeling.
Data are presented as n (%). Each number and its corresponding percentage in the criteria and category represent the count and proportion of articles that meet each specific criterion, respectively.
*The criterion was excluded from the calculation of each article's adherence rate, as it is not scorable; †Common examination approaches based on the study type: for studies involving exclusively structured data, coefficients and sensitivity analysis are often appropriate; for studies involving unstructured data in the domains of image analysis or natural language processing, saliency maps or equivalents are often appropriate.

https://doi.org/10.3349/ymj.2024.0007

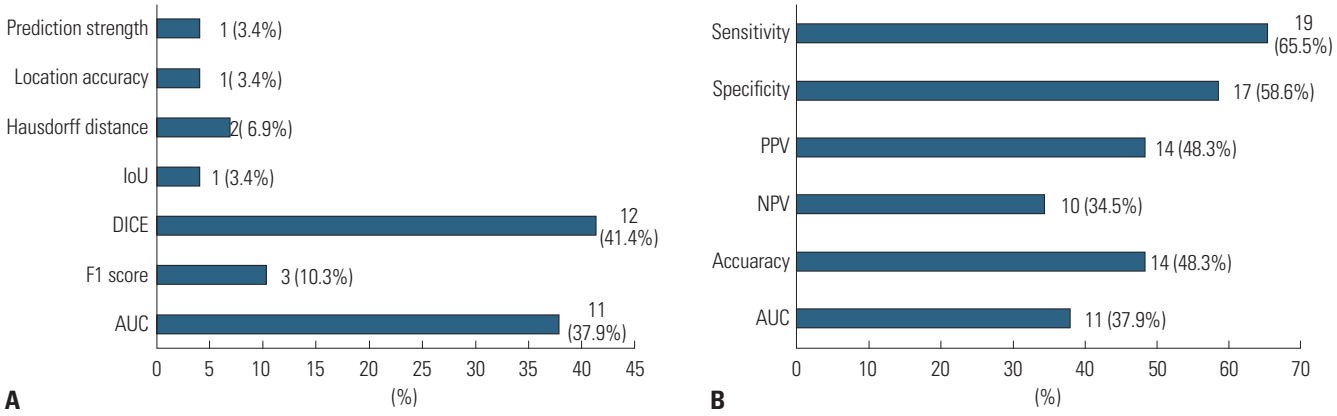

Fig. 3. Assessment of performance metrics of 29 reviewed AI/ML articles. (A) General performance metrics. (B) Clinical performance metrics. IoU, intersection over union; DICE, dice similarity coefficient; AUC, area under the curve; PPV, positive predictive value; NPV, negative predictive value; AI, artificial intelligence; ML, machine learning.

ous general performance metrics were reported with different frequencies. As shown in Fig. 3A, the dice similarity coefficient (DICE) was the most commonly reported metric featured in 12 articles, corresponding to 41.4% of the total. This was followed by the area under the curve, which was present in 11 articles, accounting for 37.9%. The F1 score was reported in 3 articles, making up 10.3% of the studies. Less frequently reported metrics included intersection over union, Hausdorff distance, and location accuracy, which were used in 1 (3.4%), 2 (6.9%), and 1 article (3.4%), respectively, reported in conjunction with DICE. Additionally, one paper that employed unsupervised learning used prediction strength as its metric. Various clinical performance metrics reported in the 29 analyzed articles are summarized in Fig. 3B. Sensitivity was the most commonly reported metric that was observed in 19 articles (65.5%), indicating its importance in the evaluation of clinical outcomes. Specificity was also frequently reported, featuring in 17 articles (58.6%), followed by accuracy and positive predictive value, each present in 14 articles (48.3%). The negative predictive value was included in 10 articles (34.5%).

Twenty-seven articles utilized unstructured data, while two articles utilized structured data. Among the 27 articles utilizing unstructured data, three included saliency maps: one was based on Gradient-weighted Class Activation Mapping (Grad-CAM), another on attention-based mechanism, and the third on CNN layer feature maps. One article suggested filtered images of specific filters, which we considered as an equivalent of saliency maps. According to the MI-CLAIM document, to fully meet the Examination Technique criteria, at least two different methods, such as saliency techniques or sensitivity analysis, must be used. In our research, only three articles satisfied this. Of the two articles featuring models using structured data, one reported using SHapley Additive exPlanations (SHAP), while the other reported beta coefficients from logistic regression for explanation purposes. However, only one article reported using at least two explanation techniques as specified in the MI-CLAIM criteria.

DISCUSSION

AI/ML studies have demonstrated promising results for various tasks related to stroke. Assessing the quality of AI/ML research is crucial for its clinical implementation. In this study, we conducted the first systematic analysis of AI/ML papers in the field of hemorrhagic stroke using the MINIMAR and MI-CLAIM checklists. Our analysis revealed significantly lower scores in certain assessment items.

In the present study, the adherence rates to MINIMAR and MI-CLAIM in various stroke-related studies were found to be 47.6% and 46.0%, respectively. These rates were consistent with the 47.4% adherence rate reported by Sohn and Won[23] for ischemic stroke-related studies assessed using MINIMAR. To be clinically applicable, AI/ML research in the field of hemorrhagic stroke requires significant improvements. In the field of AI in neuro-oncology, Kouli, et al.[24] reviewed 84 articles on automated brain tumor detection and segmentation, assessing them with the CLAIM criteria. Similar to our findings, a vast majority (95.3%) of these studies disclosed the source of their data. However, only a small number provided details on handling missing data sizes (2.6%) or the demographics of their cases (6%). There were also notable discrepancies in the reporting of model architecture (83.8%) and source code availability (24.4%), which were more comprehensive compared to our study. In the realm of AI in neurodegenerative diseases, some study groups assessed the article quality using guidelines such as QUADAS-2 and PROBAST.[25,26] Due to the challenges in applying all the criteria of QUADAS-2, they mainly reported risks of bias stemming from only some criteria of QUADAS-2.[25]

Unlike research on neuro-oncology/ neurodegenerative disease, hemorrhagic stroke studies predominantly utilized CT (89.7%). This is likely because CT is the primary modality used for early stroke assessment in clinical practice, whereas MRI is not typically performed for hemorrhagic stroke unless there are suspected underlying conditions such as vascular malformations or brain tumors.

When assessing the relationship between the adherence rates to MINIMAR/MI-CLAIM and the impact factors of the journals in which the articles were published, no significant correlation was found (Pearson's R=0.296 for MINIMAR and 0.286 for MI-CLAIM, with $p>0.05$ for both).

The MINIMAR assessment revealed a low reporting rate of patient cohort characteristics, such as race, ethnicity, and socioeconomic status. MINIMAR emphasizes the importance of standardizing the study population and patient demographics to enable fair comparisons between study outcomes. Only a few studies reported on the race and ethnicity of the cohort, and these were typically multicenter trials. Single-center AI/ML studies rarely provided this information. In terms of the model architecture component, studies rarely suggested potential users or beneficiaries of the developed model. Identifying the intended users is crucial, as it helps prevent misinterpretation of model outcomes. Only eight studies demonstrated a list of variables (27.6%).[22,27-33] In 27 studies, unstructured data were handled without explicitly presenting variables. Regarding the model evaluation component, the adherence rate was the lowest (37.9%). This was mainly due to a lack of model transparency, as most studies, except for three,[17-19] did not make their codes publicly available.

In the MI-CLAIM assessment, only two studies provided details on the evaluated models and the code developed to select the best model.[20,21] This finding was consistent with the overall observation that code transparency was low across the studies. Additionally, limited number of studies employed sufficient model examination techniques to enhance model interpretability.[22,29] Along with case-based analysis of the model performed, one study using structured data applied SHAP, while the other study dealing with unstructured data used Grad-CAM as a saliency map, discussing its feasibility and interpretability.[34,35] It is worth noting that more recent and advanced methods have been proposed, demonstrating improved performances. Considering these trends, researchers should be encouraged to actively employ such examination tools to enhance the interpretability of their models. In the present study, only 17.2% of the included studies presented a state-of-the-art model as a baseline for comparison. However, a recent systematic review of radiology AI in major subspecialties reported that performance comparison with state-of-the-art models was conducted in 37% of cases.[36] This discrepancy suggests that comparisons with state-of-the-art models may be less prevalent in the neuroradiology or stroke domains than in other subspecialties. During the analysis, we encountered some ambiguous descriptions within the checklists, leading to discussions among the reviewers. Specifically, in MI-CLAIM, the discussion of the relevance of the examination results with respect to the model/algorithm performance (component #4) was somewhat unclear. This highlights the need for enhanced precision and clarity in defining certain checklist criteria.

It is noteworthy that these two assessment tools exhibit distinct characteristics in model evaluation. MINIMAR includes a specific sub-category for external validation, which MI-CLAIM lacks. In MI-CLAIM, model performance and model examination are addressed separately, with an emphasis in the model examination category especially for explainable AI. However, these criteria are seldom met in the reviewed articles, resulting in varied adherence rates between the frameworks. Furthermore, MI-CLAIM does not include a transparency criterion in the model evaluation category, contributing to the discrepancy in adherence rates between MINIMAR and MI-CLAIM and highlighting their divergent approaches to model transparency.

The study had several limitations. The evaluation criteria we employed are not absolute benchmarks but represented just a selection of a few available evaluation tools. For example, TRIPOD, which is traditionally used for predictive model development studies and may not ideally suit AI/ML studies, has recently been updated in the form of TRIPOD-AI.[37] It would be beneficial to apply this newly published version in future research. In some instances, applying criteria designed for both AI and ML does not perfectly fit the specific model types. In MINIMAR, the substantial allocation of points to patient cohort information seems to be a disadvantage. Comparing MINIMAR's results with other AI research evaluation metrics could prove insightful. Additionally, both MINIMAR and MI-CLAIM have not yet been updated to fully encompass the features of generative models, such as large language models, vision language models, or other diffusion-based models. These frameworks need refinement to better reflect the capabilities and diverse outputs of these advanced technologies. Finally, the limited number of studies included in our analysis prevented further subdivision into meaningful subgroups. In conclusion, the present study highlights the need for enhanced interpretability and transparency in AI/ML studies on hemorrhagic stroke, as evidenced by the adherence rates to MINIMAR and MI-CLAIM. Addressing these challenges is crucial to ensure the reliability and clinical applicability of AI and ML tools in this field.

AUTHOR CONTRIBUTIONS

Conceptualization: So Yeon Won. Data curation: Chae Young Lim, Beomseok Sohn, Min Jung Seong, and So Yeon Won. Formal analysis: Chae Young Lim, Beomseok Sohn, and So Yeon Won. Investigation: Chae Young Lim, Beomseok Sohn, and So Yeon Won. Methodology: Beomseok Sohn. Project administration: So Yeon Won. Supervision: Eung Yeop Kim and Sung Tae Kim. Validation: Chae Young Lim and Beomseok Sohn. Visualization: Chae Young Lim. Writing—original draft: Chae Young Lim. Writing—review & editing: all authors. Approval of final manuscript: all authors.

ORCID iDs

Chae Young Lim https://orcid.org/0009-0000-7858-5430
Beomseok Sohn https://orcid.org/0000-0002-6765-8056

https://doi.org/10.3349/ymj.2024.0007

Minjung Seong https://orcid.org/0000-0002-9257-4225
Eung Yeop Kim https://orcid.org/0000-0002-9579-4098
Sung Tae Kim https://orcid.org/0000-0001-8185-0063
So Yeon Won https://orcid.org/0000-0003-0570-3365

REFERENCES

1. Feigin VL, Norrving B, Mensah GA. Global burden of stroke. Circ Res 2017;120:439-48.
2. GBD 2016 Stroke Collaborators. Global, regional, and national burden of stroke, 1990-2016: a systematic analysis for the global burden of disease study 2016. Lancet Neurol 2019;18:439-58.
3. van Asch CJ, Luitse MJ, Rinkel GJ, van der Tweel I, Algra A, Klijn CJ. Incidence, case fatality, and functional outcome of intracerebral haemorrhage over time, according to age, sex, and ethnic origin: a systematic review and meta-analysis. Lancet Neurol 2010;9:167-76.
4. Feigin VL, Krishnamurthi RV, Parmar P, Norrving B, Mensah GA, Bennett DA, et al. Update on the global burden of ischemic and hemorrhagic stroke in 1990-2013: the GBD 2013 study. Neuroepidemiology 2015;45:161-76.
5. An SJ, Kim TJ, Yoon BW. Epidemiology, risk factors, and clinical features of intracerebral hemorrhage: an update. J Stroke 2017;19:3-10.
6. Powers WJ, Rabinstein AA, Ackerson T, Adeoye OM, Bambakidis NC, Becker K, et al. Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. Stroke 2019;50:e344-418.
7. Hemphill JC 3rd, Greenberg SM, Anderson CS, Becker K, Bendok BR, Cushman M, et al. Guidelines for the management of spontaneous intracerebral hemorrhage: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. Stroke 2015;46:2032-60.
8. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017;14:749-62.
9. 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.
10. Bluemke DA, Moy L, Bredella MA, Ertl-Wagner BB, Fowler KJ, Goh VJ, et al. Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers—from the Radiology editorial board. Radiology 2020;294:487-9.
11. 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.
12. 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.
13. 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;26:1364-74.
14. Lekadir K, Osuala R, Gallin C, Lazrak N, Kushibar K, Tsakou G, et al. FUTURE-AI: guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. arXiv [Preprint]. 2021 [accessed on 2024 January 17]. Available at: https://doi.org/10.48550/arXiv.2109.09658.
15. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med 2020;26:1320-4.
16. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (minimum information for medical AI reporting): developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc 2020;27:2011-5.
17. Sharrock MF, Mould WA, Ali H, Hildreth M, Awad IA, Hanley DF, et al. 3D deep neural network segmentation of intracerebral hemorrhage: development and validation for clinical trials. Neuroinformatics 2021;19:403-15.
18. Muschelli J, Sweeney EM, Ullman NL, Vespa P, Hanley DF, Crainiceanu CM. PItcHPERFeCT: primary intracranial hemorrhage probability estimation using random forests on CT. Neuroimage Clin 2017;14:379-90.
19. Jianbo C, Hanqi P, Yihao C, Cheng J, Hong S, Yuxiang W, et al. Weakly supervised multitask learning models to identify symptom onset time of unclear-onset intracerebral hemorrhage. Int J Stroke 2022;17:785-92.
20. Barros RS, van der Steen WE, Boers AM, Zijlstra I, van den Berg R, El Youssoufi W, et al. Automated segmentation of subarachnoid hemorrhages with convolutional neural networks. Inform Med Unlocked 2020;19:100321.
21. Arab A, Chinda B, Medvedev G, Siu W, Guo H, Gu T, et al. A fast and fully-automated deep-learning approach for accurate hemorrhage segmentation and volume quantification in non-contrast whole-head CT. Sci Rep 2020;10:19389.
22. Yin HL, Jiang Y, Huang WJ, Li SH, Lin GW. A magnetic resonance angiography-based study comparing machine learning and clinical evaluation: screening intracranial regions associated with the hemorrhagic stroke of adult moyamoya disease. J Stroke Cerebrovasc Dis 2022;31:106382.
23. Sohn B, Won SY. Quality assessment of stroke radiomics studies: promoting clinical application. Eur J Radiol 2023;161:110752.
24. 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.
25. Pellegrini E, Ballerini L, Hernandez MDCV, Chappell FM, González-Castro V, Anblagan D, et al. Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: a systematic review. Alzheimers Dement (Amst) 2018;10:519-35.
26. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529-36.
27. Zhang L, Li J, Yin K, Jiang Z, Li T, Hu R, et al. Computed tomography angiography-based analysis of high-risk intracerebral haemorrhage patients by employing a mathematical model. BMC Bioinformatics 2019;20(Suppl 7):193.
28. Tang Z, Zhu Y, Lu X, Wu D, Fan X, Shen J, et al. Deep learning-based prediction of hematoma expansion using a single brain computed tomographic slice in patients with spontaneous intracerebral hemorrhages. World Neurosurg 2022;165:e128-36.
29. Rangaraj S, Islam M, Vs V, Wijethilake N, Uppal U, See AAQ, et al. Identifying risk factors of intracerebral hemorrhage stability using explainable attention model. Med Biol Eng Comput 2022;60:337-48.
30. Li H, Xie Y, Liu H, Wang X. Non-contrast CT-based radiomics score for predicting hematoma enlargement in spontaneous intracerebral hemorrhage. Clin Neuroradiol 2022;32:517-28.
31. Seymour SE, Rava RA, Swetz DJ, Monteiro A, Baig A, Schultz K, et al. Predicting hematoma expansion after spontaneous intracranial hemorrhage through a radiomics based model. Proc SPIE Int Soc Opt Eng 2022;12033:120332X.

32. Skoch J, Tahir R, Abruzzo T, Taylor JM, Zuccarello M, Vadivelu S. Predicting symptomatic cerebral vasospasm after aneurysmal subarachnoid hemorrhage with an artificial neural network in a pediatric population. Childs Nerv Syst 2017;33:2153-7.
33. Liu J, Xu H, Chen Q, Zhang T, Sheng W, Huang Q, et al. Prediction of hematoma expansion in spontaneous intracerebral hemorrhage using support vector machine. EBioMedicine 2019;43:454-9.
34. Fu R, Hu Q, Dong X, Guo Y, Gao Y, Li B. Axiom-based Grad-CAM: towards accurate visualization and explanation of CNNs. arXiv [Preprint]. 2020 [accessed on 2024 January 17]. Available at: https://doi.org/10.48550/arXiv.2008.02312.
35. Jung H, Oh Y. Towards better explanations of class activation mapping [accessed on 2024 January 17]. Available at: https://doi.org/10.1109/ICCV48922.2021.00137.
36. Kelly BS, Judge C, Bollard SM, Clifford SM, Healy GM, Aziz A, et al. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). Eur Radiol 2022;32:7998-8007.
37. Cohen JF, Bossuyt PMM. TRIPOD+AI: an updated reporting guideline for clinical prediction models. BMJ 2024;385:q824.

https://doi.org/10.3349/ymj.2024.0007