AJNR

Evaluating Biases and Quality Issues in Intermodality Image Translation Studies for Neuroradiology: A Systematic Review

Shannon L. Walston, Hiroyuki Tatekawa, Hirotaka Takita, Yukio Miki and Daiju Ueda

AJNR Am J Neuroradiol 2024, 45 (6) 826-832
doi: https://doi.org/10.3174/ajnr.A8211
http://www.ajnr.org/content/45/6/826

Evaluating Biases and Quality Issues in Intermodality Image Translation Studies for Neuroradiology: A Systematic Review

Shannon L. Walston, Hiroyuki Tatekawa, Hirotaka Takita, Yukio Miki, and Daiju Ueda

ABSTRACT

BACKGROUND: Intermodality image-to-image translation is an artificial intelligence technique for generating one technique from another.

PURPOSE: This review was designed to systematically identify and quantify biases and quality issues preventing validation and clinical application of artificial intelligence models for intermodality image-to-image translation of brain imaging.

DATA SOURCES: PubMed, Scopus, and IEEE Xplore were searched through August 2, 2023, for artificial intelligence–based image translation models of radiologic brain images.

STUDY SELECTION: This review collected 102 works published between April 2017 and August 2023.

DATA ANALYSIS: Eligible studies were evaluated for quality using the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) and for bias using the Prediction model Risk Of Bias ASsessment Tool (PROBAST). Medically-focused article adherence was compared with that of engineering-focused articles overall with the Mann-Whitney *U* test and for each criterion using the Fisher exact test.

DATA SYNTHESIS: Median adherence to the relevant CLAIM criteria was 69% and 38% for PROBAST questions. CLAIM adherence was lower for engineering-focused articles compared with medically-focused articles (65% versus 73%, $P < .001$). Engineering-focused studies had higher adherence for model description criteria, and medically-focused studies had higher adherence for data set and evaluation descriptions.

LIMITATIONS: Our review is limited by the study design and model heterogeneity.

CONCLUSIONS: Nearly all studies revealed critical issues preventing clinical application, with engineering-focused studies showing higher adherence for the technical model description but significantly lower overall adherence than medically-focused studies. The pursuit of clinical application requires collaboration from both fields to improve reporting.

ABBREVIATION: AI = artificial intelligence

Artificial intelligence (AI)-based image translation converts an image into a similar-but-different image.[1,2] This feature may mean changing a landscape from a summer to a winter scene, or a CT into an MR image. The accuracy and capacity to do what humans physically cannot has always been the promise of AI in medicine.[3] In neuroradiology specifically, using an AI model to convert among radiologic image modalities such as PET, MR imaging, and CT has several advantages, including increased accessibility and decreased time and radiation exposure. In the case of MR imaging, for example, patients with metal implants or contrast allergies cannot undergo the examination, though they may be able to undergo a CT.[4] An AI model produces the image almost immediately, while scheduling the examination can take days; this issue is known to affect prognosis.[5] For radiography, CT, and PET, patients are exposed to ionizing radiation, and the repeat examinations required for radiation therapy can cause cumulative damage. Thus, image translation models have been in development to harness these advantages for MR imaging-only radiation therapy planning[6] and ischemic stroke lesion segmentation since 2017.[7-10]

Despite these potential advantages and a 6-year history of published research, intermodality image translation models are

Received November 21, 2023; accepted after revision January 27, 2024.

From the Department of Diagnostic and Interventional Radiology (S.L.W., H.Tatekawa, H.Takita, Y.M., D.U.), Graduate School of Medicine, and Smart Life Science Lab (D.U.), Center for Health Science Innovation, Osaka Metropolitan University, Osaka, Japan.

Please address correspondence to Daiju Ueda, MD, PhD, Department of Diagnostic and Interventional Radiology, Osaka Metropolitan University Graduate School of Medicine, 1-4-3 Asahi-machi, Abeno-ku, Osaka 545-8585, Japan; e-mail: ai.labo.ocu@gmail.com

Indicates article with online supplemental data.

http://dx.doi.org/10.3174/ajnr.A8211

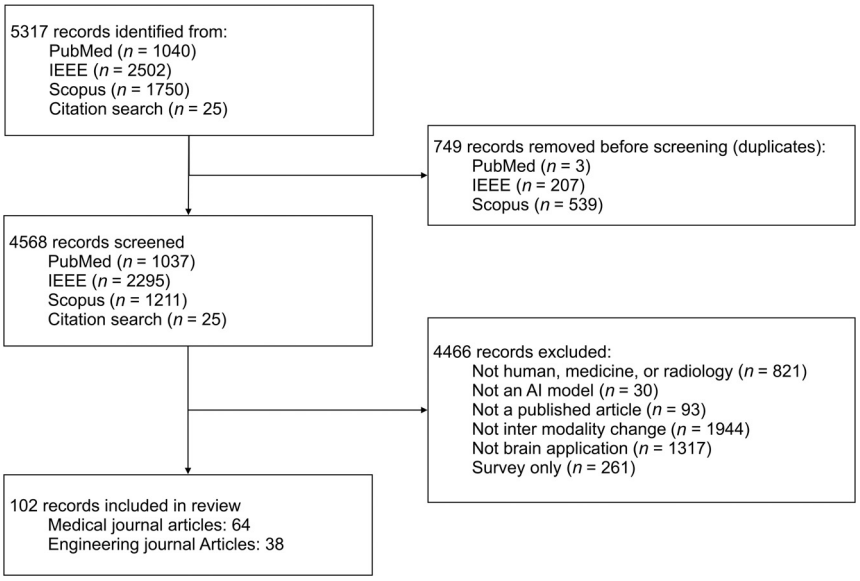

FIG 1. PRISMA flow chart.

still in the initial development stage. It has been suggested that fewer than one-quarter of AI studies could be reproduced from these methods,[11] reproduction being a necessary validation step before clinical application. Previous reviews described some study-design trends that might be related to the lack of progress toward clinical application.[4,6,12,13] There are also trends in the results for medically-focused journals and engineering-focused journals.[13,14] Various checklists have been designed to support authors in this task, but few are applicable to AI, and even fewer, to imaging-based studies. The Checklist for Artificial Intelligence in Medical Imaging (CLAIM)[15] is a prominent checklist for AI models built to classify, generate, or otherwise use medical images. This checklist includes 42 items that authors should include to ensure that readers can thoroughly assess and reproduce the work. Additionally, the Prediction model Risk Of Bias ASsessment Tool (PROBAST)[16] is designed for assessing bias in diagnostic or prognostic prediction models. This bias assessment is an integral part of any systematic review of health care models because it allows readers to visualize which studies have shortcomings that may lead to distorted results. Although many questions are not applicable to AI models, some PROBAST items can be used to evaluate biases specific to image-to-image translation studies. Using multiple checklists may provide more comprehensive coverage of all the salient points of each work.[17]

As this field grows, researchers must be aware of and consider the quality of and biases in their methods so that they can be transparently and consistently reported and, eventually, systematically mitigated.[18] In this review we used 2 common checklists to quantify the quality and extent of biases in intermodality image translation studies for brain imaging. We found no study applying CLAIM or PROBAST to evaluate image-to-image translation articles in the field of brain imaging.

MATERIALS AND METHODS

This review was registered on PROSPERO (CRD42022368642; https://www.crd.york.ac.uk/PROSPERO/) and was conducted in accordance with the Prisma statement (https://www.prisma.io/).[19] Approval from the ethics board was not necessary because this review used published data.

Searching Strategy

PubMed, Scopus, and IEEE Xplore were searched from inception through August 2, 2023, using variations of the terms artificial intelligence, MR, CT, image-to-image translation, image synthesis, Pix2Pix, and GAN. The full search text is available in the Online Supplemental Data. After a preliminary search, keywords related to the brain or specific brain diseases were considered too limiting, so we excluded these terms to reduce sampling bias. The references of similar reviews and the included studies were also screened for inclusion. Duplicate results were removed, and the remaining published articles were independently screened for inclusion by 2 authors. Inclusion criteria were studies developing or evaluating an AI model capable of converting images of the brain from one image technique to another, from human participants. Only the relevant experiments were considered from studies with multiple experiments (Fig 1).

Data

The data required to identify each study were collected into a predesigned spreadsheet. This spreadsheet includes relevant article information, data set information, the overall model purpose and design, the translation pair for all relevant experiments, and results for all relevant CLAIM and PROBAST criteria.[15,16] To reveal any differences specific to articles published in medically-focused or engineering-focused journals, we grouped journals on the basis of their aims and scope as in Kim et al.[12] Medically-focused journals were defined as those that included terms related to clinical medicine in their scope, and engineering-focused journals were those with an engineering, physics, or computer science scope. Unclear journals were classified in consensus between 2 authors. Extractable data were collected by one author and confirmed by 3 authors.

Quality Evaluation

Adherence to the CLAIM checklist was evaluated by 1 author using the full text and supplements of each study. Values were considered absent if they were missing or unclear. Questions not relevant to the study were marked as not applicable and did not negatively affect the CLAIM adherence estimation. For example, not all studies were classification or diagnosis tasks, so CLAIM question 36 was not applicable to these studies (Online Supplemental Data).

Bias Evaluation

PROBAST-based bias was evaluated by 1 author using questions 1.1, 1.2, 4.1, and 4.8 from PROBAST.[16] We used a generous definition of "external data set" for question 4.8, which includes

Table 1: Included studies

	Number of Studies	Average CLAIM Adherence	Average PROBAST Score
Image-generation direction			
MR imaging-CT	63	71%	38%
MR imaging-PET	13	74%	49%
CT-MR imaging	12	63%	41%
PET-CT	3	67%	33%
PET-MR	2	74%	56%
MR imaging-x-ray	1	64%	31%
US-MR imaging	1	67%	31%
Bidirectional	7	65%	33%
Total	102	70%	39%

Note:—US indicates ultrasound.

temporally separate data as well as data from different facilities, as in Kim et al.[12] Following Kuo et al[20] and Nagendran, et al,[21] we considered the patients included in the test set for domain 1 and excluded the other questions as not relevant to AI models, which perform image-to-image translation. This includes all of domain 2, which is not relevant to AI studies, and all of domain 3, because the outcome is not relevant for image translation studies (Online Supplemental Data).

For paired-image studies, which developed diagnostic, prognostic, or segmentation models, a large gap between the imaging of the 2 ground truth modalities may affect the resultant model classification or segmentation performance. The appropriate timing depends on the specific disease and may vary for individuals within a data set, so this information was collected but not evaluated.

Analysis

For CLAIM, the number of items evaluated as "yes" or "not applicable" was summed and divided by the total items in the checklist to estimate adherence at the study level as in Sivanesan et al.[14] PROBAST adherence is given as "high," "unclear," or "low" risk of bias designations for each study. Data normality was assessed using the Shapiro-Wilk test. Means were compared using the 2-sample *t* test; medians were compared using the Mann-Whitney *U* test between articles in medically-focused publications and those from engineering-focused publications. Item-level evaluation was performed for both CLAIM and PROBAST to show trends in image-to-image translation research. The Fisher exact test was used for each criterion to compare the medically-focused studies with the engineering-focused studies. Significance was defined as $P < .05$. Analysis was performed using R (Version 4.1.3; http://www.r-project.org/).

RESULTS

Study Demographics

There were 102 studies collected for this review. Medically-focused publications made up 64 studies, and 38 were from engineering-focused journals. The source images included MR imaging, CT, PET, and radiography (Table 1). Most studies evaluated MR imaging translation to either CT (63/102) or PET (13/102), citing the better soft-tissue contrast and lack of radiation exposure of MR imaging. Most MR imaging-to-CT studies targeted MR imaging–based radiation therapy planning (48/63). Dosimetry evaluations were included in 17 of these studies. Other targets included attenuation correction (10/63), more accurate registration (3/63), segmentation (1/63), and research data set generation for future AI studies (1/63). The MR imaging-to-PET studies primarily focused on the diagnosis of Alzheimer disease (4/13) or MS (1/13), glioma management and prognosis (2/13), attenuation correction (1/13,) and amyloid-burden estimation (2/13).

CT-to-MR imaging translation (12/102) was the next most common, with the rationale being that CT data are useful for dose calculations of radiation therapy and can be collected quickly, at lower cost, and at more facilities than MR imaging. Most of these studies targeted anomaly segmentation such as ischemic stroke localization (6/12) or CT-based radiation therapy planning (3/12). There were 7 studies that evaluated bidirectional translations; 6 of which comprised the MR imaging-to-CT and CT-to-MR imaging translation pair and 1 study that evaluated the PET-to-CT and CT-to-PET translation as well. One of these studies was designed for segmentation of follow-up images of patients with ischemic stroke; 2 others were designed to assist with radiation therapy planning.

Other translation pairs included MR imaging-to-radiograph for interventional imaging, PET-to-CT for attenuation correction, PET-to-MR imaging for amyloid-burden estimation, and ultrasound-to-MR imaging for easier communication between technicians and obstetricians (Online Supplemental Data).

The specific clinical purpose or application for the model in addition to image generation was described in 95 studies. We determined the remaining studies using the common clinical purposes for the translation pair. These include diagnosis, prognosis, registration, segmentation, and treatment (Online Supplemental Data). There was a significant difference in the CLAIM scores between the medically-focused and engineering-focused groups for both diagnosis and treatment purposes (Online Supplemental Data). The other purpose groups did not have enough data for the Mann-Whitney *U* test.

CLAIM Evaluation

Each study followed between 44% and 88% of applicable CLAIM criteria, with a 70% average overall (Fig 2). There was a significant difference in the adherence between medically-focused journal studies (73% average adherence) and those from engineering-focused journals (65% average adherence) ($P < .001$) (Fig 2 and Table 2). There was no significant difference between the performance of studies published before or after the CLAIM criteria were published. ($P = .841$).

There were 5 criteria with an average adherence of ≤10%. Of note, 1% of studies described the intended sample size (CLAIM 19), 3% of studies described the flow of including participants (CLAIM 33), and 8% of studies used an external testing data set (CLAIM 32).

Engineering-focused studies reported data for the questions related to describing the model in adequate detail (CLAIM 22, 24, 25, 26) in the article, while medically-focused studies more often moved this information to the supplement. On the other hand, medically-focused studies significantly outperformed

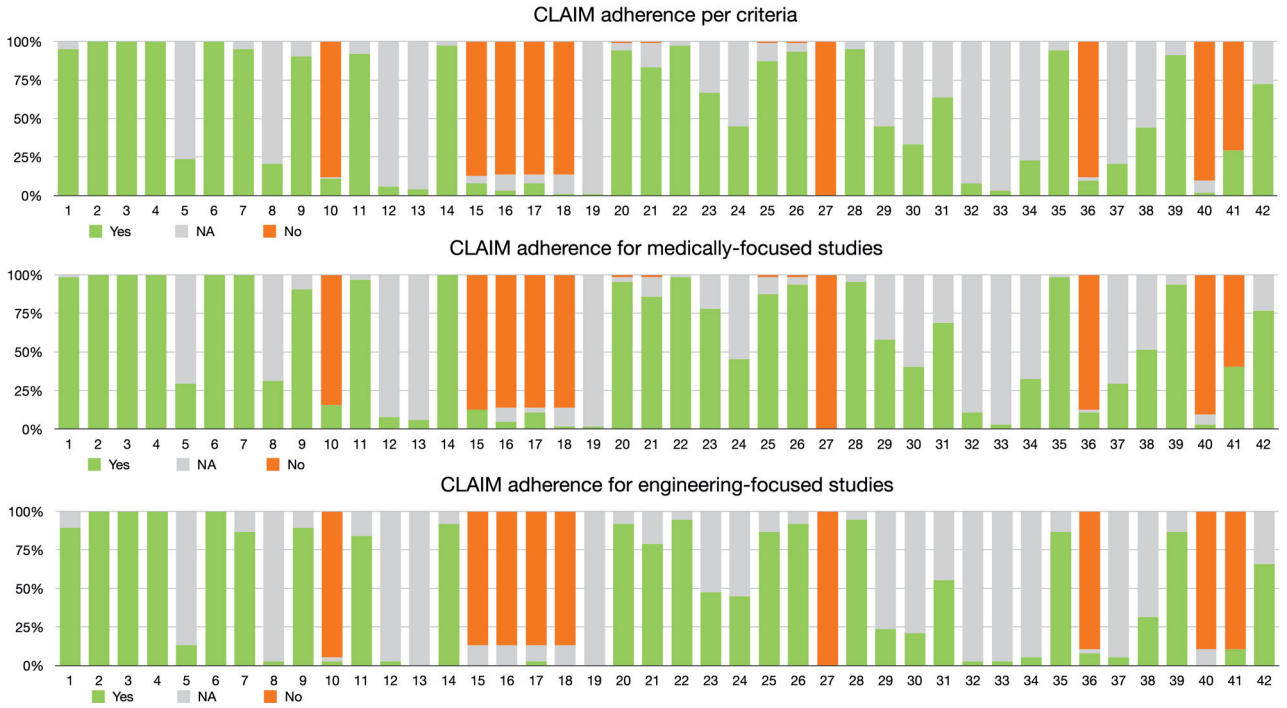

FIG 2. CLAIM evaluation. Each *vertical bar* shows the adherence for all studies for one of the CLAIM criteria. Within each bar, green represents the percentage of studies appropriately adhering to the CLAIM criteria, gray represents studies for which that question was not applicable, and orange represents studies that did not adhere to that CLAIM criteria. *A*, Overall adherence for 102 studies. *B*, Adherence for medically-focused studies ($n = 64$). *C*, Adherence for engineering-focused studies ($n = 38$). NA indicates not applicable.

Table 2: CLAIM adherence results

	Medically-Focused	Engineering-Focused	*P* Value
Title/abstract	99%	95%	.0652
Introduction	100%	100%	1
Methods	72%	63%	<.001
Results	53%	39%	.0046
Discussion	73%	59%	.0629
Other	90%	85%	.2023
Total	73%	65%	<.001

engineering-focused studies in 21% (9/42) of the CLAIM criteria. These criteria related to describing the data sets (CLAIM 7, 8, 11, 14, 15, 34), the software used for model development (CLAIM 23), statistical significance levels (CLAIM 29), and any failures (CLAIM 37).

Publication styles differ for engineering studies, and this feature significantly affected the resultant CLAIM adherence. Specifically, 27 of the 39 engineering-focused works were presented at engineering conferences and published as proceedings. These are given a DOI and widely regarded as citable publications,[22] though the peer review and writing of these submissions may still be of lower quality.[23,24] Thus, conference publications represent 72% of engineering-focused studies but <5% of medically-focused studies. Although they are considered published, we found that works written for engineering-focused conferences had significantly lower resultant CLAIM adherence than engineering-focused works for journals ($P = .01$) (Online Supplemental Data). Without these conference publications, medically-focused studies significantly outperform the engineering-focused studies for only 3 criteria related to describing the data sets (CLAIM 5, 8, 34).

The adherence for medically-focused and engineering-focused studies varied by study purpose (Online Supplemental Data). Medically-focused studies had significantly higher adherence for MR imaging-only radiation therapy planning, with an 11% improvement over the average for engineering-focused studies. Only medically-focused studies attempted dose calculations as part of a radiation therapy planning study, and excluding these did not affect the significance of the difference between the CLAIM score of these and the engineering-focused studies. Although there were not enough studies to confidently establish significance, attenuation correction and stroke lesion localization also had 10% and 18% higher CLAIM adherence than similar engineering-focused studies, respectively.

PROBAST Evaluation

Overall bias via PROBAST was low for 4 studies,[8,25-27] unclear for 4 studies, and high for 94 studies (Fig 3 and Table 3). Medically-focused studies used population-based data in significantly more studies. ($P = .006$) There was no significant difference in the PROBAST adherence for the other 3 questions between studies from medically-focused and engineering-focused journals (Fig 3) (Online Supplemental Data). There was no significant difference between the performance of studies published before or after the PROBAST criteria were published. ($P = 1$).

More than 71% (73/102) of studies used internally collected data for the test data set. These were presumed to be consecutive samples and marked as probably having a low risk of bias for question 1.1 unless stated otherwise. The remaining 29 studies used publicly available data. Although the use of curated public data sets is generally considered appropriate for AI model

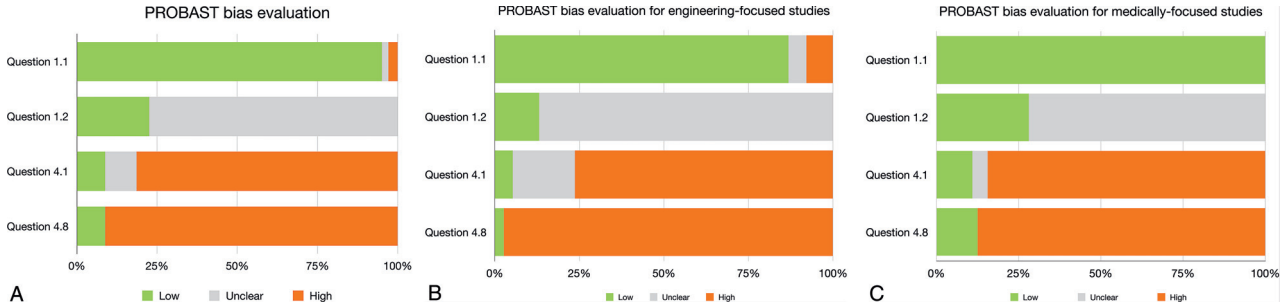

FIG 3. Bias evaluation. Each *horizontal bar* shows the risk of bias for all studies for one of the PROBAST criteria. Within each bar, green represents the percentage of studies with a low risk of bias and gray represents studies for which there was an unclear risk of bias, and orange represents studies with a high risk of bias for the question. Question 1.1 asks if the data source matched the target population. Question 1.2 asks if the inclusion and exclusion criteria were appropriate. Question 4.1 asks if the test data set was appropriately sized. Question 4.8 asks if the model was tested on an external data set to account for overfitting or optimism in the model. *A*, Overall adherence per question for all 102 studies. *B*, Adherence for engineering-focused studies ($n = 38$). *C*, Adherence for medically-focused studies ($n = 64$).

Table 3: Bias risk results

	Low	Unclear	High
All studies			
Question 1.1	97 (95%)	2 (2%)	3 (3%)
Question 1.2	23 (23%)	79 (77%)	0 (0%)
Question 4.1	9 (9%)	10 (10%)	83 (81%)
Question 4.8	9 (9%)	0 (0%)	93 (91%)
Overall bias	4 (4%)	4 (4%)	94 (92%)
Medically-focused adherence			
Question 1.1	64 (100%)	0 (0%)	0 (0%)
Question 1.2	18 (28%)	46 (72%)	0 (0%)
Question 4.1	7 (11%)	3 (5%)	54 (84%)
Question 4.8	8 (13%)	0 (0%)	56 (88%)
Overall bias	3 (5%)	4 (6%)	55 (86%)
Engineering-focused adherence			
Question 1.1	33 (87%)	2 (5%)	3 (8%)
Question 1.2	5 (13%)	33 (87%)	0 (0%)
Question 4.1	2 (5%)	7 (18%)	29 (77%)
Question 4.8	1 (3%)	0 (0%)	37 (97%)
Overall bias	1 (3%)	0 (0%)	37 (97%)

development, PROBAST standards require the model to be tested on data representative of the target population. Studies using public data sets that were not collected or adjusted to match the sampling frequencies of a real population were marked as having a high risk of bias for question 1.1.

Exclusion criteria were described in 17 studies, 13 of which were from medically-focused journals. An additional 6 studies (5 of which were from medically-focused journals) included language that implied that there were some exclusion criteria, though they were not described in detail. These were marked as probably having a low risk of bias for question 1.2. Because it is unclear if exclusions were maliciously hidden or if the authors had included all available images, the remaining 79 studies were marked as having an unclear risk of bias.

There were 6 studies with >100 individuals in the test data set, as recommended by PROBAST. Five of these were published in a medically-focused journal. The remaining studies either had test sets of <100 individuals or were not clear about the number of images of individuals in their test data set (this point is further addressed in CLAIM 21).

External data sets (PROBAST 4.8 and CLAIM 32) were used in 9% (9/102) of studies. Eight of these were published in medically-focused journals. The size of the external test set varied with a median of 17.5 and interquartile range of 126.25. It is difficult to accurately calculate significant changes across time with only 9 studies.

The timing between the index and reference test was reported in 30 of 89 studies that used paired data, and 3 of 13 studies that used unpaired data. We observed a slightly higher reporting rate in medically-focused studies (38% [24/64]) compared with engineering-focused studies (24% [9/38]) ($P = .191$).

DISCUSSION

This systematic review evaluated the quality issues and biases present in intermodality image translation studies relevant to brain imaging published before August 2023 using PROBAST and CLAIM criteria. We found 102 studies using brain images for intermodality image translation using an AI model. The principal findings of this review are that nearly all of the 102 published works had quality issues and critical biases hindering the clinical integration progress, with engineering-focused studies showing significantly lower checklist adherence than medically-focused studies. Studies at a high risk of bias largely lacked an external testing data set and were unclear about the data used, particularly the collection dates, data-source location, any exclusion criteria, the number of individuals included, and how the data were processed to make them fit for the AI model. Replication of results is of great importance in medicine—reflected in the numerous, detailed checklists available to researchers—and replication in the age of AI requires closer collaboration with our computer engineering colleagues.

To our knowledge, this is the first study to evaluate study quality and biases in intermodality image-to-image translation models for brain imaging. CLAIM and PROBAST were used to evaluate whether the methods or data sets used in these studies showed quality issues or risks of bias.[15,16] PROBAST is not only familiar to many readers, but it covers 4 of the signs of bias raised in other reviews and guidelines for applying AI to medical imaging translation.[12,13,28] We additionally chose the CLAIM checklist because it is designed to show the "rigor, quality, and generalizability of the work," by encouraging transparent and thorough reporting specifically of medical AI studies. This

checklist not only addresses all 4 included PROBAST criteria but also exposes bias risks from inadequate reporting of the included data sets and model development methods. The use of these measures together gives us a more granular view of the strengths and weaknesses of these studies.[17]

Because AI-based neurologic image translation represents an intersection of the medical and engineering fields, researchers on both sides must work together to make sure their work is clearly represented in published articles to improve repeatability.[11] Engineering studies in journals effectively described 4 CLAIM questions about the AI model (CLAIM 22, 24–26) more often than medically-focused studies, though this difference was not statistically significant. Similarly, medically-focused studies more often included data set details such as the number of patients and their demographics, included statistical measures such as confidence intervals, and listed some known limitations of their work (CLAIM 5, 8, 29, 34, 38). For these models to continue toward clinical integration, they must prove good performance while following both medical and AI engineering standards. Medically-focused tools such as PROBAST and CLAIM may be improved by requiring more of the model details provided by AI engineers to ensure accurate replication of these ever more complex AI models.[29] Additionally, authors can introduce checklists such as CLAIM for imaging studies, STARD (https://www.equator-network.org/reporting-guidelines/stard/)[30] for diagnostic accuracy studies, and TRIPOD (https://www.tripod-statement.org/about/)[31] for diagnostic or prognostic prediction studies to their collaborators who may be unfamiliar with them as a guide for writing their sections. Open collaboration between medical and AI engineering researchers is key to moving these models past the initial development stage.

Our findings of extensive risks of bias do not imply that a validated AI model would be insufficient as a clinician's support tool. Current tools and procedures have weaknesses for which AI may be able to compensate.[32] For example, there are scenarios such as radiation therapy planning that benefit from having both the better soft-tissue contrast of MR imaging and the electron density estimates of CT.[33] Because it is not always practical or possible to perform both examinations, an image translation model could generate this image. Furthermore, the CT scans and MR images must be registered for tasks like atlas-based methods or ischemic stroke lesion localization, which can lead to artifacts from misalignment.[4,6,34] By negating these weaknesses, image-generation AI models can lead to speedier, safer, more cost-efficient workflows, benefitting both the patient and the facility.

Limitations

This review had several limitations. There was heterogeneity in study designs of the collected studies, limiting our ability to compare the models and data sets directly. Our inclusion criteria may have excluded relevant studies, though we attempted to correct this possibility by scanning the references of both the collected works and previous reviews on the topic. We did not compare the publication requirements for the included journals, so it is unclear whether requirements, such as word-count limits, supplementary-material limits, or requirements on the use of standardized checklists, impacted quality and bias estimates. This study used only 2 checklists to estimate study design failures and risk of bias. While PROBAST is a common tool for bias evaluation of medical studies, it is difficult to apply to AI and even more so for image translation models. New guidelines specific to medical AI applications are in development that may address this difficulty.[29,35,36] Perhaps future AI works will modify their methods accordingly to minimize bias.

CONCLUSIONS

Image-to-image translation AI models represent a promising tool for reducing radiation exposure, examination costs, and time delay. However, currently published models have quality issues and are at high risk of bias, attributable to weak adherence to established reporting guidelines such as CLAIM and PROBAST. From a clinical applicability point of view, studies published in engineering-focused journals have significantly more quality issues and higher risk of bias than those published in medically-focused journals. However, medically-focused studies often lack necessary model development details found in engineering-focused studies. Our analysis shows that closer cooperation between medical and engineering researchers could improve overall guideline adherence, so these models can be validated and developed into valuable clinical tools.

Disclosure forms provided by the authors are available with the full text and PDF of this article at www.ajnr.org.

REFERENCES

1. Wolterink JM, Mukhopadhyay A, Leiner T, et al. **Generative adversarial networks: a primer for radiologists.** *Radiographics* 2021;41:840–57 CrossRef Medline
2. Ueda D, Shimazaki A, Miki Y. **Technical and clinical overview of deep learning in radiology.** *Jpn J Radiol* 2019;37:15–33 CrossRef Medline
3. Matheny ME, Whicher D, Thadaney Israni S. **Artificial Intelligence in Health Care: A Report From the National Academy of Medicine.** *JAMA* 2020;323:509–10 CrossRef Medline
4. Laino ME, Cancian P, Politi LS, et al. **Generative adversarial networks in brain imaging: a narrative review.** *J Imaging* 2022;8:83 CrossRef Medline
5. Seymour ZA, Fogh SE, Westcott SK, et al. **Interval from imaging to treatment delivery in the radiation surgery age: how long is too long?** *Int J Radiat Oncol Biol Phys* 2015;93:126–32 CrossRef Medline
6. Sorin V, Barash Y, Konen E, et al. **Creating artificial images for radiology applications using generative adversarial networks (GANs): a systematic review.** *Acad Radiol* 2020;27:1175–85 CrossRef Medline
7. Feng E, Qin P, Chai R, et al. **MRI generated from CT for acute ischemic stroke combining radiomics and generative adversarial networks.** *IEEE J Biomed Health Inform* 2022;26:6047–57 CrossRef Medline
8. Garzon G, Gomez S, Mantilla D, et al. **A deep CT to MRI unpaired translation that preserve ischemic stroke lesions.** *Conf Proc IEEE Eng Med Biol Soc* 2022;2022:2708–11 CrossRef Medline
9. Gutierrez A, Tuladhar A, Wilms M, et al. **Lesion-preserving unpaired image-to-image translation between MRI and CT from ischemic stroke patients.** *Int J Comput Assist Radiol Surg* 2023;18:827–36 CrossRef Medline
10. Rubin J, Abulnaga SM. **CT-to-MR conditional generative adversarial networks for ischemic stroke lesion segmentation.** In: *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, June 10–13, 2019. Xi'an, China CrossRef
11. Bontempi D, Nuernberg L, Krishnaswamy D, et al. **Transparent and Reproducible AI-based Medical Imaging Pipelines Using the**

Cloud. *Research Square*. https://doi.org/10.21203/rs.3.rs-3142996/v1. Accessed October 1, 2023
12. Kim DW, Jang HY, Kim KW, et al. **Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers.** *Korean J Radiol* 2019;20:405–10 CrossRef Medline
13. England JR, Cheng PM. **Artificial intelligence for medical image analysis: a guide for authors and reviewers.** *AJR Am J Roentgenol* 2019;212:513–19 CrossRef Medline
14. Sivanesan U, Wu K, McInnes MD, et al. **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;74:334–42 CrossRef Medline
15. Mongan J, Moy L, Kahn CE Jr. **Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers.** *Radiol Artif Intell* 2020;2:e200029 CrossRef Medline
16. Wolff RF, Moons KG, Riley RD, et al; PROBAST Group. **PROBAST: a tool to assess the risk of bias and applicability of prediction model studies.** *Ann Intern Med* 2019;170:51–58 CrossRef Medline
17. Klontzas ME, Gatti AA, Tejani AS, et al. **AI reporting guidelines: how to select the best one for your research.** *RadiolArtif Intell* 2023;5:e230055 CrossRef Medline
18. Ueda D, Kakinuma T, Fujita S, et al. **Fairness of artificial intelligence in healthcare: review and recommendations.** *Jpn J Radiol* 2024;42:3–15 CrossRef Medline
19. Page MJ, Moher D, Bossuyt PM, et al. **PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews.** *BMJ* 2021;372:n160 CrossRef Medline
20. Kuo RY, Harrison C, Curran T-A, et al. **Artificial intelligence in fracture detection: a systematic review and meta-analysis.** *Radiology* 2022;304:50–62 CrossRef Medline
21. Nagendran M, Chen Y, Lovejoy CA, et al. **Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies.** *BMJ* 2020;368:m689 CrossRef Medline
22. Patterson D, Snyder L, Ullman J. **Evaluating computer scientists and engineers for promotion and tenure.** *Computer Research Association* 1999 Aug 1. [Epub ahead of print]
23. Vardi MY. **Conferences vs. journals in computing research.** *Communications of ACM* 2009;52:5 CrossRef
24. Fortnow L. **Viewpoint: Time for computer science to grow up.** *Communications of ACM* 2009;52:33–35 CrossRef
25. Choi H, Lee DS, Alzheimer's Disease Neuroimaging Initiative. **Generation of structural MR images from amyloid PET: application to MR-less quantification**. *J Nucl Med* 2018;59:1111–17 CrossRef Medline
26. Pan Y, Liu M, Lian C, et al. **Synthesizing missing PET from MRI with cycle-consistent generative adversarial networks for Alzheimer's disease diagnosis**. *Med Image Comput Comput Assist Interv* 2018; 11072:455–63 CrossRef Medline
27. Takita H, Matsumoto T, Tatekawa H, et al. **AI-based virtual synthesis of methionine PET from contrast-enhanced MRI: development and external validation study.** *Radiology* 2023;308:e223016 CrossRef Medline
28. Park SH, Han K. **Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction.** *Radiology* 2018;286:800–09 CrossRef Medline
29. Collins GS, Dhiman P, Andaur Navarro CL, 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 CrossRef Medline
30. Bossuyt PM, Reitsma JB, Bruns DE, et al; STARD Group. **STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies.** *BMJ* 2015;351:h5527 CrossRef Medline
31. Moons KG, Altman DG, Reitsma JB, et al; Transparent Reporting of a Multivariate Prediction Model for Individual Prognosis or Development Initiative. **New guideline for the reporting of studies developing, validating, or updating a multivariable clinical prediction model: the tripod statement.** *Adv Anat Pathol* 2015;22:303–05 CrossRef Medline
32. Willemink MJ, Koszek WA, Hardell C, et al. **Preparing medical imaging data for machine learning.** *Radiology* 2020;295:4–15 CrossRef Medline
33. Wang T, Lei Y, Fu Y, et al. **A review on medical imaging synthesis using deep learning and its clinical applications.** *J Appl Clin Med Phys* 2021;22:11–36 CrossRef Medline
34. Yonezawa H, Ueda D, Yamamoto A, et al. **Maskless 2-dimensional digital subtraction angiography generation model for abdominal vasculature using deep learning.** *J Vasc Interv Radiol* 2022;33:845–51.e8 CrossRef Medline
35. Tejani AS, Klontzas ME, Gatti AA, et al. **Updating the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) for reporting AI research.** *Nat Mach Intell* 2023;5:950–51 CrossRef
36. Sounderajah V, Ashrafian H, Golub RM, et al; STARD-AI Steering Committee. **Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol.** *BMJ Open* 2021;11:e047709 CrossRef Medline